New Statistical Learning Methods for Evaluating Dynamic Treatment Regimes and Optimal Dosing

by

Ming Tang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2020

Doctoral Committee:

Professor Jeremy M.G. Taylor, Co-Chair
Professor Lu Wang, Co-Chair
Professor Brian Denton
Research Professor Matthew J. Schipper

Ming Tang

mingtang@umich.edu

ORCID iD: 0000-0002-6451-8648

To my parents and Wenkang.

ACKNOWLEDGEMENTS

This dissertation is a summary of my past six-year research and academic experience at the University of Michigan. It is written under the fears and tension of the COVID pandemic and the civil rights movements. The chaos has taken a heavy toll on health, life, suffering, and security. Under such chaotic circumstances, I feel like being educated in a decent research institute is a privilege that many people may dream of. The department of Biostatistics of the University of Michigan has been a safe and warm home in the United States for me and also nourishes me with scholarship and knowledge. I am so fortunate to finish my dissertation with the guidance and encouragement from advisors, friends, family, and collaborators. You share this amazing journey with me, so it only seems right that I dedicate this dissertation to all of you.

First of all, I am grateful to my advisor Lu Wang for her constant encouragement and insightful academic advice— she has been a role model for me since I first started my Ph.D. study. Her critical insights into the research questions as a statistician shape my academic knowledge base and improve my understanding as a scholar. As an international student away from my family, I also enjoy sharing my happiness and frustration with her because she is so considerate and supportive. If I become only half the scientist, half the statistician, half the person that Lu is, it will surely be one of my greatest accomplishments.

I want to express my gratitude to my advisor Jeremy M.G. Taylor. I have been studying from him since he was the instructor of BIOSTAT699. He is a respected

scholar who is very keen on the methods we studied and detail-oriented with the dataset we worked with. With his wisdom, guidance, support, and patience, Jeremy has taught me to be a rigorous statistician.

I owe my special thanks to my GSRA advisor Matthew Schipper for supporting my Ph.D. study and advising my GSRA work. Matt taught me skills in working efficiently with collaborators and connecting statistical insights into medical research problems. My second project is a perfect practice inspired by our collaboration. This project will not be possible without his inspiration.

I am also grateful to Brian Denton for his generosity in offering the application dataset of my first project and providing practical inputs and considerations. Yilun Sun is another person who had an important academic influence on me. I have gained a lot of hands-on experience by working with the third project with him.

I also want to thank Ivo Dinov from the Statistical Online Computing Resource for supporting my master's study and encouraging me to pursue a Ph.D. degree. Special thanks go to Cathie Spino, who was my "supervisor" when I worked at SOCR. She is a great mentor in encouraging me under difficulties with her appropriate career advice. Also, she has been a role model for me as a scientist but also as a strong woman when I first knew her life experience.

I am indebted to my research collaborators from the Michigan Medicine Radiation Oncology Department. Thanks go to Daniel Spratt, William Jackson, Dawn Owen, Ray Dipankar, Corey Speers, and Chris Maurino. The working experience of the cutting-edge clinical oncology projects with them helps me prepare my future career in the pharmaceutical industry. Very special thanks are to Ted Lawerence. He has been a renowned medical expert but still eagers to learn pioneering statistical methods from junior statisticians, like me, to help him better treat patients. His work ethic towards dedication motivates me profoundly.

Last but not least, I would like to thank my parents Jianxi Huang and Jiangbo

Tang for their endless encouragement. I also want to express my gratitude and love to my husband, Wenkang Huang. No matter if I am at my good or bad times, he always listens to me and never stops supporting me pursuing my academic goal, even in such a long-distance relationship. I am so lucky to have him in my life.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Dynamic treatment regimes (DTRs) have gained increasing interest in the field of personalized health care in the last two decades, as they provide a sequence of individualized decision rules for treating patients over time. In a DTR, treatment is adapted in response to the changes in an individual's disease progression and health care history.

However, specific challenges emerge when applying the current methods of DTR in practice. For example, a treatment decision often happens after a medical test, and is thus nested within the decision of whether a test is needed or not. Such nested test-and-treat strategies are attractive to improve cost-effectiveness. In the first project of this dissertation, we develop a Step-adjusted Tree-based Learning (SAT-Learning) method to estimate the optimal DTR within such a step-nested multiple-stage multiple-treatment dynamic decision framework using test-and-treat observational data. At each step within each stage, we combine a doubly robust semiparametric estimator via Augmented Inverse Probability Weighting with a tree-based reinforcement learning procedure to achieve the counterfactual optimization. SAT-Learning is robust and easy to interpret for the strategies of disease screening and subsequent treatments when necessary. We applied our method to a Johns Hopkins University prostate cancer active surveillance dataset to evaluate the necessity of prostate biopsy and identify the optimal test-and-treatment regimes for prostate cancer patients.

Our second project is motivated by scenarios in medical practice where one need to decide on patients radiation or drug doses over time. Due to the complexity of continuous dose scales, few existing studies have extended their methods of multi-treatment decision making to a method to estimate the optimal DTR with continuous doses. We develop a new method, Kernel-Involved-Dosage-Decision learning (KIDD-Learning), which combines a kernel estimation of the dose-response function with a tree-based dose-search algorithm, in a multiple-stage setting. At each stage, KIDD-Learning recursively estimates a personalized dose-response function using kernel regression and then identifies the interpretable optimal dosage regime by growing an interpretable decision tree. The application of KIDD-Learning is illustrated by evaluating the dynamic dosage regimes of the adaptive radiation therapy using a Michigan Medicine liver cancer dataset.

In KIDD-Learning, our algorithm splits each node of a tree-based decision rule from the root node to terminal nodes. This heuristic algorithm may fail to identify the optimal decision rule when there are critical tailoring variables hidden from an imperceptible parent node. Therefore, in the third project, we propose an important modification of KIDD-Learning, Stochastic Spline-Involved Tree Search (SSITS), to estimate a more robust optimal dosage regime. This new method uses a simulated annealing algorithm to stochastically search the space of tree-based decision rules. In each visited decision rule, a non-parametric smooth coefficient model is applied to estimate the dose-response function. We further implement backward induction to estimate the optimal regime from the final stage in a reverse sequential order to previous treatment stages. We apply SSITS to determine the optimal dosing strategy for patients treated with Warfarin using data from the International Warfarin Pharmacogenetics Consortium.

CHAPTER I

Introduction

Personal health care has gained increasing attention in recent decades (*Collins and Varmus*, 2015; *Niculescu et al.*, 2019). Compared with the traditional "one-size-fits-all" paradigm, this medical approach allows physicians to consider the heterogeneity of patients and yields more precise treatment recommendations and better disease management. (*Shi et al.*, 2020)

In particular, when managing a chronic disease, a personalized disease management plan may involve multiple cycles of treatments, and the treatment in each cycle is chosen adaptively based on patients' history and time-varying characteristics (*Lee et al.*, 2015). This paradigm is known as a Dynamic Treatment Regime (DTR) (*Murphy*, 2003; *Wang et al.*, 2012), in which the entire sequence of decisions is evaluated, rather than evaluating each treatment separately. Identifying optimal DTRs offers an effective tool for personalized management of diseases, and helps physicians tailor the treatment strategies dynamically and individually based on clinical evidence, which provides a key foundation for enhanced care of chronic disease. (*Murphy*, 2003; *Chakraborty and Murphy*, 2014).

However, it is challenging to develop a method for identifying the optimal dynamic treatment regime due to the complex relationship between the clinical outcomes and the accumulated time-varying treatment information. The time-varying confounders

should be carefully adjusted because the standard regression methods may fail to address this complexity. Some pioneering statistical methods have been developed to estimate the optimal DTRs, such as the G-estimation of structural nested mean models (*Robins*, 1986, 1989), the Marginal Structural Model estimated with the inverse probability weighting (*Robins et al.*, 2000; *Robins*, 2004), the Marginal Mean Model (*Murphy et al.*, 2001; *Murphy*, 2003), and other likelihood-based methods (*Thall et al.*, 2007). These methods often require a parametric or semi-parametric conditional model for the counterfactual outcome as a component and thus can be vulnerable to model mis-specification, especially when the data are high dimensional.

More recently, as an alternative for parametric or semi-parametric models, machine learning-based approaches have become increasingly popular because of their flexibility in model assumptions and their robustness against model misspecification. Examples include Q-learning (*Watkins and Dayan*, 1992), A-learning (*Murphy*, 2003; *Schulte et al.*, 2014), and the backward outcome weighted learning (*Zhao et al.*, 2015; *Chen et al.*, 2018), all of which use backward induction to first optimize the decision rule for the final stage, and then optimize the decisions of the previous stages in a reversed sequential order.

However, specific research questions in clinical practice cannot be addressed when applying the current methods of DTR. Therefore, in this dissertation, we aim to continue this research direction and develop three robust and interpretable statistical learning methods to estimate the optimal DTR for various types of decisions in multiple-stage settings. In particular, we use the counterfactual framework of causal inference (*Robins*, 1986), and estimate the optimal DTR by maximizing the expectation of the counterfactual outcome. Furthermore, due to the need applying these innovative statistical learning methods to medical practice, a robust and interpretable method for estimating the optimal DTR is highly desirable, as it bridges the gap between the physician's medical expertise and the data-driven individualized treatment

regimes, and allows a physician to better understand and apply them.

Our first project as presented in Chapter II considers handling a special but important test-and-treat decision strategy in the application of DTR. It is motivated by the example of the active surveillance of prostate cancer (*Lange et al.*, 2018; *Tosoian et al.*, 2015). For prostate cancer patients, a decision about treatment happens after the biopsy, and is thus nested within the decision of whether to do the biopsy test. Considering the substantial side-effects of over-treatment and unnecessary biopsy screening, the active surveillance, which involves closely watching patients' disease conditions but not giving any treatment unless significant progression is found, has been increasingly recommended to patients with low risk prostate cancer (*Denton et al.*, 2019). However, the one-size-fits-all active surveillance plan is not individualized for patients with heterogeneity. For example an older male in good health might tolerate the invasive biopsy test, while a younger man might not be appropriate to be monitored under the same active surveillance schedule because of his comorbidities (*Loeb et al.*, 2014). A personalized sequence of an active surveillance plan, i.e., the biopsy test and definitive treatment plan, is a paradigm that actually resembles a DTR, as it accounts for patients' time-varying characteristics. However, the current DTR methods cannot accommodate such a naturally embedded property of the treatment decision within the test decision. Therefore, we developed a new statistical learning method to evaluate DTRs within such a nested multi-stage dynamic decision framework using observational data. At each step within each stage, we combined the robust semi-parametric estimator via Augmented Inverse Probability Weighting with a tree-based reinforcement learning method to deal with the counterfactual optimization. The proposed method can handle test-and-treat observational data and estimate an interpretable and robust optimal DTR for the strategies of disease screening and subsequent treatments if necessary.

The second (Chapter III) and third project (Chapter IV) focus on estimating the

optimal DTR with continuous treatment options in a multiple-stage setting. This research direction is motivated by the study of Michigan Medicine liver cancer adaptive stereotactic body radiation therapy (SBRT). In this study, the liver cancer patients were treated with adaptive SBRT with an intra-treatment evaluation (*Feng et al.*, 2018). The previous data analysis of this study has shown the benefit of this innovative adaptive SBRT in terms of improving overall survival probability and controlling the incidence of local progression (*Feng et al.*, 2013). Physicians are now interested in the optimal personalized dose schema for a future trial under this adaptive SBRT framework. That is, what dose should be given for patients for the first stage, and should patients terminate the SBRT after the first stage due to high toxicity? If the treatment continues, what is the radiation dose for the second stage, based on the patients' individualized time-varying characteristics?

How to estimate the optimal dosage regime within a multiple-stage setting is a common research question in oncology when treating cancer patients with radiation or drug doses over time. However, it is complex to extend the existing methods for multiple-treatment to personalized dose-finding because of the sparse nature of the observed data, i.e., the dose level follows a continuous distribution – of which the probability of observing a specific dose is zero (*Chen et al.*, 2016). Although considerable research has been devoted to estimating multiple-stage DTRs with multiple treatments in the literature, less attention has been paid to the estimation of DTRs with continuous treatment options (*Lee et al.*, 2015; *Chen et al.*, 2016; *Rich et al.*, 2016; *Schulz and Moodie*, 2020). In Chapter III and Chapter IV, we develop two methods, Kernel-Involved-Dose-Decision Learning (KIDD-Learning) and Stochastic Spline-Involved Tree Search (SSITS), to estimate the optimal DTR with continuous doses. In particular, in KIDD-Learning, a non-parametric kernel regression is utilized to estimate a robust continuous dose-response function while in SSITS a flexible smooth coefficient model is applied to evaluate the continuous dose effect. Both of

the non-parametric methods are combined with flexible dose-search statistical learning methods to identify the optimal continuous dosage regime in the setting of a multiple-stage DTR.

In addition to the robust non-parametric estimation of the dose-response function, the choice of dose-search methods of SSITS and KIDD-Learning is another critical contribution of this dissertation. Since an interpretable treatment regime is straightforward for physicians to understand, we believe a tree-based decision rule is an appropriate choice for the desired interpretability. Therefore in the second project (Chapter III), we use a tree-based dose-search algorithm to identify the optimal dosing strategy. This algorithm, which is derived from the Classification and Regression Tree (CART) (*Breiman et al.*, 1984; *Laber and Zhao*, 2015), yields a tree-based dose decision rule by categorizing patients into different sub-population based on their characteristics. When combining the algorithm with the estimation of the dose-response function as the KIDD-Learning method, the satisfactory performance is illustrated by comprehensive simulation studies .

However, from the simulations studies, we found that in some cases, when there are strong predictors hidden from a relatively weak parent node, KIDD-Learning may end up estimating a sub-optimal tree-based decision rule. This limitation is derived from the heuristic nature of the CART, which is used in the dose-search algorithm in KIDD-Learning. Such a top-down algorithm generates a sequence of trees, each of which is a direct extension of the previous decision tree, which will inevitably result in a locally optimal tree if the parent node is an imperceptible variable. Therefore, in Chapter IV, we propose an improvement of KIDD-Learning based on a stochastic dose-search algorithm, SSITS, for estimating the optimal dosage regime. Rather than searching heuristically for the optimal dose from the root node to the terminals, SSITS stochastically visits a broader binary decision tree space via the simulated annealing algorithm and then determines the optimal dosing strategy. Compared to its CART

counterpart, SSITS can efficiently search tree space more widely to escape from a local optimal decision rule while still delivering the optimal DTR with satisfactory interpretability. The outstanding stable performance of SSITS is well demonstrated in the simulation studies.

CHAPTER II

Step-adjusted Tree-based Reinforcement Learning for Evaluating Nested Dynamic Treatment Regimes with Test-and-Treat Observational Data

2.1 Introduction

Dynamic treatment regimes (DTRs) have gained increasing interest in the field of precision medicine in the last decade (*Chakraborty and Murphy*, 2014). This research direction generalizes the individualized medical decisions into a time-varying treatment setting, usually at discrete stages, and thus accommodates the updated information for each person at each stage (*Murphy*, 2003; *Wang et al.*, 2012). In DTR, actions or decisions based on the individualized features are able to lead to more precise disease prevention and better disease management. However, the current DTR framework is limited because it only considers choosing the best treatments strategies. In medical practice, the procedures to diagnose and treat patients are much more complicated. Most diagnosis procedures or tests, e.g., positron emission tomography, or a biopsy test, occur prior to the selection of treatment to provide more information about disease status, then this information would be used to select treatment. Typically, only patients who have taken the test can be treated, and thus the decision about the treatment assignment is nested within the decision of performing

the test.

For example, men with early stage asymptomatic prostate cancer who are in an active surveillance program, would regularly have their prostate-specific antigen (PSA) and prostate tissue measured via a blood test and core needle biopsy test respectively (*Loeb et al.*, 2014). Whether to undergo definitive treatment for their prostate cancer would be strongly influenced by the results from their biopsy test. So the possible treatment initiation only happens after having the biopsy test result, and is thus nested within the decision of doing a biopsy or not. Such a nested dynamic clinical decision-making is not limited to prostate cancer. The occult blood test, also known as a stool test, can also be used as a cheap and easy initial screening test for colorectal cancer (*Itzkowitz et al.*, 2008). Patients with abnormal finding from the stool test are then referred for a colonoscopy exam, which is costly and invasive, to confirm the diagnosis and decide if more definitive treatment for colorectal cancer is needed. In this scenario the decision of whether to do definitive treatment is nested in the decision of whether to do a colonoscopy which is nested within the decision to do a stool test or not. This kind of nested clinical decision also happens with many other chronic diseases. (*US Preventive Services Task Force*, 2009; *Mandelblatt et al.*, 2009; *Hanley*, 2011)

In such nested test-and-treat scenarios, the impact of the test should also be considered. For some diseases, the tests used to confirm the diagnosis or decide on the next step are easy to administer and minimally invasive, e.g., blood test and physical examination. But some other tests done for confirmatory purposes are expensive and invasive, including the prostate biopsy and colonoscopy. The potential side effects include pain, soreness, and infections, which should not be overlooked. For prostate cancer, even if the test result suggests progressive disease, it is not always the case that the patient should undergo definitive treatment, which has substantial comorbidity, since prostate cancer is a slow growing disease and a substantial number

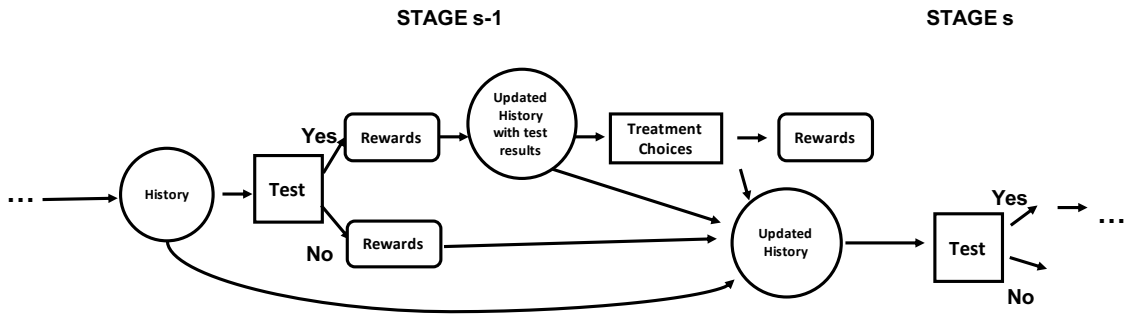

Figure 2.1: Hypothetical step-adjusted DTR framework with a treatment step nested within the test step of each intervention stage.

of men may not develop deadly prostate cancer before dying from some other cause. It is well known that there is overtreatment for prostate cancer, and that a substantial number of men receive unnecessary cancer treatments (*Loeb et al.*, 2014). Therefore, careful patient selection for testing is needed to not only reduce the impact on the patient, but also to save medical resources for the patients who truly need them. The current one-fits-all active surveillance protocol is not capable of taking the patient's personalized medical characteristics into account and then giving an individualized disease management plan.

As mentioned above, most existing frameworks for evaluating DTRs overlook such nested structures during the clinical decision making process. The diagnostic test itself does not have a direct impact on the disease-related outcome, but the potential treatment following the test may improve the disease outcome for the patient substantially (*Trikalinos et al.*, 2009). Overlooking such a test-and-treat nested structure may result in identifying imprecise and non-realistic decision rules. Therefore, we propose a new nested dynamic treatment regime (nested-DTR) framework *by embedding the treatment step within the test step of each intervention stage* as shown in Figure 2.1. At each stage, the decision of the test step is made based on the health history and the treatment decision is made on the basis of previous health history and the updated history after the test.

In general, DTRs can be estimated from observational data, provided there is enough heterogeneity in the patient features and their actions taken. Similarly the optimal DTR for this new nested-DTR framework can be learned from observational data provided there is enough heterogeneity in data for both the decision to test and the decision to treat.

A great number of statistical methods have been developed to estimate the optimal DTRs using observational data, such as Marginal Structural Model estimated with inverse probability weighting (*Robins*, 2004), the Marginal Mean Model (*Murphy et al.*, 2001) and other likelihood-based methods (*Thall et al.*, 2007). These methods require a parametric or semi-parametric conditional model for the counterfactual outcome as a component and thus are vulnerable to model mis-specification, especially when the data are high dimensional or time-dependent information is accumulated. More recently, machine learning-based approaches, as a replacement for parametric or semi-parametric models, have become increasingly popular because of their flexibility in model assumptions and their robustness (*Laber et al.*, 2014; *Zhao et al.*, 2015). When identifying the optimal DTRs with multiple stages, the problem resembles the reinforcement learning (RL) problem (*Watkins and Dayan*, 1992). Therefore, RL methods are currently broadly applied in evaluating the optimal DTRs. Some of this work, which involving reinforcement learning, has focused on developing easily interpretable DTRs for real-world practice (*Shen et al.*, 2017; *Tao and Wang*, 2017; *Tao et al.*, 2018; *Schulte et al.*, 2014; *Zhang et al.*, 2015).

To the best of our knowledge, however, none of the existing methods can be applied directly to estimate the optimal DTRs when each stage consists of a treatment step nested within a test step. In this paper, we are trying to fill this gap and develop a new statistical learning method for identifying the optimal DTR within the nested dynamic decision framework. At each step within each stage, we combine the robust semi-parametric estimator obtained using Augmented Inverse Probability Weighting

(AIPW) with a modified tree-based reinforcement learning method to optimize the expected counterfactual outcome. The remainder of this paper is organized as follows: In 2.2 and 2.3, we formalize the problem of identifying the optimal DTR within the nested DTR framework in a multiple-stage multiple-step setting from observational data and develop the nested step-adjusted tree-based reinforcement learning method (SAT-Learning). Section 4 presents the detailed implementation of this new method. Numerical simulation studies and an application to the Johns Hopkins University (JHU) prostate cancer active surveillance data are provided in 2.5 and 2.6. We conclude with a brief discussion in Section 2.7.

2.2 Multi-Stage Nested Step-Adjusted Dynamic Treatment Regimes

To address the nested decision problem above, we consider a nested multi-stage multi-step decision framework with S decision stages. In clinical practice, every regular clinic visit, which might initiate some form of treatment, can be considered as a stage. Within each stage s, there are J action steps. Let K_{sj} denote the number of decision options at step j of stage s ($K_{sj} \geq 2$), let D_{sj} denote the multiple treatment indicators of the action taken at step j of stage s in the observed data, and the value of D_{sj} is $d_{sj} \in \mathcal{D}_{sj}$. Without loss of generality, we consider two steps within each stage, i.e., $J = 2$, to make the presentation easier. We assume the first step of stage s is the test step (action D_{s1}) and D_{s2} in the treatment step is nested within the decision of D_{s1}. For example, only the prostate cancer patients who have had the biopsy test are considered for further treatment. We denote the patient's history prior to action D_{sj} but after the previous step as X_{sj}. We will use overbar with subscripts s and j to denote a vector of a variables's history up to the step j of stage s. For example, $\overline{X}_{s2} = (X_{11}, X_{12}, X_{21}, \ldots, X_{s1}, X_{s2})$. Similarly, the action history up to the

treatment step of stage s can be denoted as $\overline{D}_{s2} = (D_{11}, D_{12}, D_{21}, \ldots, D_{s1})$.

We use Y_{sj} to denote the intermediate reward outcome at the end of step j of the stage s, and thus the overall rewards vector is $(Y_{11}, Y_{12}, \ldots Y_{S2})$. The outcome of interest Y is a function of all rewards, i.e., $Y = f(Y_{11}, Y_{12}, Y_{21}, \ldots Y_{S2})$, where $f(\cdot)$ is a pre-specified function (e.g., sum). We also assume that Y is bounded and high values of Y are desirable. The observed data before stage s step j $(1 \leq s \leq S, 1 \leq j \leq 2)$ are

$$\{X_{11}, D_{11}, Y_{11}, X_{12}, \ldots, D_{s-1,2}, Y_{s-1,2}, X_{s1}\}_{i=1}^{n} \equiv \{\overline{X}_{s1}, \overline{D}_{s-1,2}, \overline{Y}_{s-1,2}\}_{i=1}^{n}$$

for step 1 , and

$$\{X_{11}, D_{11}, Y_{11}, X_{12}, \ldots, X_{s1}, D_{s1}, Y_{s1}, X_{s2}\}_{i=1}^{n} \equiv \{\overline{X}_{s2}, \overline{D}_{s1}, \overline{Y}_{s1}\}_{i=1}^{n}$$

for step 2. For brevity, we suppress the subject index i in the following text when no confusion exists. The observed data are assumed to be independent and identically distributed for n subject from the population of interest. The history $\mathbf{H}_{sj}$ is defined as the test results and action history prior to the action assignment D_{sj}. To be more specific, $\mathbf{H}_{s1} = (\overline{D}_{s-1,2}, \overline{X}_{s1}, \overline{Y}_{s-1,2})$ and $\mathbf{H}_{s2} = (\overline{D}_{s1}, \overline{X}_{s2}, \overline{Y}_{s1})$. To illustrate the method, we also specify two action options in the test step and three options in the treatment step of every stage, i.e., $d_{s1} \in \mathcal{D}_{s1} = \{0, 1\}$, $K_{s1} = 2$, and $d_{s2} \in \mathcal{D}_{s2} = \{0, 1, 2\}$, $K_{s2} = 3$. When a patient has $d_{sj} = 0$, i.e. no treatment or test is given, he/she will still be kept in the study cohort but not given further treatment until the next stage $s + 1$. Thus, the reward is $Y_{sj} = 0$ when $d_{sj} = 0$. For the data collected from the active surveillance study , if the patient receives treatment in some treatment step, i.e., $d_{s2} = 1$ or 2, he will be removed from the study according to the active surveillance protocol.

With a treatment step nested after every test step within a stage, the nested DTR

is defined as a personalized test-and-treatment rule sequence. The rule is based on the observed history $\mathbf{H}_{sj}$ about the patient's health status up to the action in step j of stage s. Let $\mathbf{g}$ denote the above nested DTR. Formally, $\mathbf{g} = (g_{11}, g_{12}, \ldots, g_{S2})$ is defined by a collection of mapping functions, where g_{sj} is mapped from the domain of history $\mathbf{H}_{sj}$ to the domain of D_{sj}, i.e.,

$$\mathbf{H}_{sj} \mapsto g_{sj}(\mathbf{H}_{sj}) \in \mathcal{D}_{sj}, 1 \leq s < S, 1 \leq j \leq 2$$

2.3 Step-adjusted Optimization for Nested DTR

Let $Y^*(\mathbf{g})$ be the counterfactual outcome if all patients follow $\mathbf{g}$ to assign treatment or test conditional on previous history. The performance of $\mathbf{g}$ is measured by the counterfactual mean outcome $E\{Y^*(\mathbf{g})\}$ conditional on the patients' history. We denote the optimal regime as $\mathbf{g}^{opt}$. Our goal of identifying the optimal regime is to find the $\mathbf{g}^{opt}$ which satisfies

$$E\{Y^*(\mathbf{g}^{opt})\} \geq E\{Y^*(\mathbf{g})\}$$

for all $\mathbf{g} \in \mathcal{G}$, where $\mathcal{G}$ is the set of all potential regimes.

2.3.1 Optimization of g_{S2} and g_{S1} for the Final Stage S

The approach to finding optimal DTR includes backward induction (*Murphy et al.*, 2001), therefore we illustrate the mathematical formulation from the last stage S. For the last step of the stage, let $Y^*_{S2}(d_{S2})$ be the counterfactual outcome if a patient makes treatment decision d_{S2} conditional on previous history. We denote the optimal regime as g^{opt}_{S2}, which satisfies $E\{Y^*_{S2}(g^{opt}_{S2})\} \geq E\{Y^*_{S2}(g_{S2})\}$ for all $g_{S2} \in \mathcal{G}_{S2}$, where $\mathcal{G}_{S2}$ is the

set of all potential regimes at stage S and step 2.

To connect the counterfactual outcome with observed data $\{\overline{X}_{S2}, \overline{D}_{S2}, \overline{Y}_{S2}\}$, we make the following standard causal inference assumptions (*Murphy*, 2003; *Orellana et al.*, 2010):

1. *Consistency.* The observed outcome coincides with the counterfactual outcome under the treatment a patient is actually given, i.e.,

$$Y_{S2} = \sum_{d_{S2} \in \mathcal{D}_{S2}} Y^*_{S2}(d_{S2}) I\{g_{S2}(\mathbf{H}_{S2}) = d_{S2}\} I\{d_{S1} = 1\},$$

   where $I(\cdot)$ is the indicator function that takes the value 1 if $\cdot$ is true and 0 otherwise. The indicator function $I(d_{S1} = 1)$ implies only the subjects who decided to take the previous test, i.e., $d_{S1} = 1$, can have their Y_{S2} observed.

2. *No unmeasured confounding.* The observed action D_{S2} is independent of potential counterfactual outcomes conditional on the history $\mathbf{H}_{S2}$, i.e.,

$$D_{S2} \perp \{Y^*_{S2}(0), Y^*_{S2}(1), Y^*_{S2}(2)\} \mid \mathbf{H}_{S2},$$

   where $\perp$ denotes statistical independence. This assumption implies that the potential confounders are fully observed and included in the dataset.

3. *Positivity.* For the observational data, the propensity score $\pi_{d_{S2}}(\mathbf{H}_{S2})$, the probability of receiving a certain treatment conditional on history, is bounded away from 0 and 1, i.e., $\pi_{d_{S2}}(\mathbf{H}_{S2}) = Pr(D_{S2} = d_{S2} \mid \mathbf{H}_{S2}) \in [c_1, c_2]$, where $0 < c_1 < c_2 < 1$.

For the subjects who do not have the test in the previous step, i.e., $d_{S1} = 0$, their test result that the further treatment decision is based on cannot be observed. Therefore, only the subjects with $d_{S1} = 1$ is able to contribute to the optimization of g_{S2}. Under the three assumptions, the optimization problem for the treatment of the last stage

becomes

$$g_{S2}^{opt}(\mathbf{H}_{S2}) = arg \max_{g_{S2}\in\mathcal{G}_{S2}} E_{\mathbf{H}_{S2}}\Big(\sum_{d_{S2}\in\mathcal{D}_{S2}} E(Y_{S2} \mid D_{S2} = d_{S2}, \mathbf{H}_{S2}) \\ I[g_{S2}(\mathbf{H}_{S2}) = d_{S2}]I(d_{S1} = 1)\Big), \tag{2.1}$$

where $E_{\mathbf{H}_{S2}}(.)$ denotes the expectation with respect to the marginal joint distribution of the observed history $\mathbf{H}_{S2}$. To derive the optimal g_{S1}^{opt} for whether to take the test, i.e., one step before the treatment step within the same stage S, we utilize the backwards induction (*Murphy*, 2003). In addition to the counterfactual outcome of stage s step j Y_{sj}^* defined in the last section, we also define a nested step-adjusted future optimized counterfactual outcome $\widetilde{Y}_{S1}^*$. More specifically, we have $\widetilde{Y}_{S1}^* = \{Y^*(\overline{D}_{S-1,2}, g_{S1}, g_{S2}^{opt})\}$, where the treatment for stage S step 2 has been optimized. To determine the optimal g_{S1}^{opt}, we propose to maximize the expected nested step-adjusted future optimized counterfactual outcome $\widetilde{Y}_{S1}^*$, i.e., $g_{S1}^{opt} = \arg\max_{g_{S1}\in\mathcal{G}_{S1}} E_{\mathbf{H}_{S1}}[\{Y^*(\overline{D}_{S-1,2}, g_{S1}, g_{S2}^{opt})\}]$.

Similarly, we assume *No Unmeasured Confounding*, $D_{S1} \perp \{\widetilde{Y}_{S1}^*(0), \widetilde{Y}_{S1}^*(1)\} \mid \mathbf{H}_{s1}$, if $d_{S1} = 1$, and if $d_{S1} = 0$, $D_{S1} \perp \{Y_{S1}^*(0), Y_{S1}^*(1)\} \mid \mathbf{H}_{s1}$; *Positivity* $\pi_{d_{S1}}(\mathbf{H}_{S1}) = Pr(D_{S1} = d_{S1} \mid \mathbf{H}_{sj}) \in [c_1, c_2]$, where $0 < c_1 < c_2 < 1$; and then the optimization problem of stage S step 1 can be written as

$$g_{S1}^{opt} = \arg\max_{g_{S1}\in\mathcal{G}_{S1}} E_{\mathbf{H}_{S1}}\big[\sum_{d_{S1}\in\mathcal{D}_{S1}} \big\{E[\widetilde{Y}_{S1}^* \mid D_{S1} = d_{S1}, \mathbf{H}_{S1}]I(d_{S1} = 1) \\ + E[Y_{S1}^* \mid D_{S1} = d_{S1}, \mathbf{H}_{S1}]I(d_{S1} = 0)\big\} I\{g_{S1}(\mathbf{H}_{S1}) = d_{S1}\}\big]. \tag{2.2}$$

Different from Eqn (2.1), the optimization process Eqn (2.2) of g_{S1}^{opt} is conducted within all eligible subjects, while the optimization of g_{S2}^{opt} is conducted only within the patients who have the test at the previous step. Although the whole cohort contributes to the optimization step in Eqn (2.2), $\widetilde{Y}_{S1}^*$ or Y_{S1}^* used in Eqn (2.2) actually depends on the test decision, i.e., d_{S1}. The subjects who had the test, i.e., $d_{S1} = 1$,

essentially have one more chance to optimize their rewards through stage S step 2 compared to those without test, and this chance is nested within the positive exam decision within the same stage.

2.3.2 Optimization of g_{s2} and g_{s1} for Any Stage Before S

For the steps of stage s before the last stage ($1 \leq s < S$), the optimal regime g_{s1}^{opt} and g_{s2}^{opt} is expressed via backward induction as well. $\widetilde{Y}_{sj}^{*}$ is defined as the nested step-adjusted future optimized counterfactual reward, which is given that all future stages' and steps' actions are already optimized. More specifically, we have $\widetilde{Y}_{s1}^{*}(g_{s1}) = \{Y^{*}(\overline{D}_{s-1,2}, g_{s1}, g_{s2}^{opt}, \ldots, g_{S2}^{opt})\}$ and $\widetilde{Y}_{s2}^{*}(g_{s2}) = \{Y^{*}(\overline{D}_{s1}, g_{s2}, g_{s+1,1}^{opt}, \ldots, g_{S2}^{opt})\}$. Similar to the assumptions for the last stage, we assume *No Unmeasured Confounding* and *Positivity*. Under these assumptions, the optimization problems at stage s step j can be written as

$$g_{s1}^{opt} = \arg\max_{g_{s1} \in \mathcal{G}_{s1}} E_{\mathbf{H}_{s1}}\Big[\sum_{d_{s1} \in \mathcal{D}_{s1}} E[\widetilde{Y}_{s1}^{*} \mid D_{s1} = d_{s1}, \mathbf{H}_{s1}] I\{g_{s1}(\mathbf{H}_{s1}) = d_{s1}\}\Big] \tag{2.3}$$

and

$$g_{s2}^{opt} = \arg\max_{g_{s2} \in \mathcal{G}_{s2}} E_{\mathbf{H}_{s2}}\Big[\sum_{d_{s2} \in \mathcal{D}_{s2}} E[\widetilde{Y}_{s2}^{*} \mid D_{s2} = d_{s2}, \mathbf{H}_{s2}] I\{g_{s2}(\mathbf{H}_{s2}) = d_{s2}\} I(d_{s1} = 1)\Big]. \tag{2.4}$$

2.4 Step-adjusted Tree-based Reinforcement Learning and its Implementations

Given the observational data with test-and-treat nested decision structure, we propose to solve Eqn (2.1), Eqn (2.2), Eqn (2.3), and Eqn (2.4) through the step-adjusted tree-based learning (SAT-Learning) method. In this method, the step-adjusted future optimized pseudo-outcome is iteratively inducted backwards. We further assume, for

stages and steps before the last step, i.e., for any $s < S$, $j = 1$ or 2,, the effect of intermediate outcome reward Y_{sj} will be cumulatively carried forward to the final outcome (*Huang et al.*, 2015), and denote a nested step-adjusted future optimized pseudo-outcome of stage s step j as PO_{sj}. Let $\mu_{sj,d_{sj}}(\mathbf{H}_{sj}) = \widehat{E}[PO_{sj} \mid D_{sj} = d_{sj}, \mathbf{H}_{sj}]$ be the estimated mean pseudo-outcome of stage s step j. Because of the cumulative property of the reward outcome and the nested connection between the test step and the treatment step, for any $s < S$, $j = 1$ or 2, PO_{sj} can be expressed in a recursive form as $PO_{s1} = Y_{s1} + \sum_{r=s}^{S} \mu_{r2,g_{r2}^{opt}}(\mathbf{H}_{r2}) \times I(d_{r1} = 1) + \sum_{r=s+1}^{S} \mu_{r1,g_{r1}^{opt}}(\mathbf{H}_{r1})$ and $PO_{s2} = Y_{s1} + \sum_{r=s+1}^{S}[\mu_{r2,g_{r2}^{opt}}(\mathbf{H}_{r2}) \times I(d_{r1} = 1) + \mu_{r1,g_{r1}^{opt}}(\mathbf{H}_{r1})]$. Obviously, when evaluating the pseudo-outcome in last stage, we have $PO_{S2} = Y_{S2}$ for the second step and $PO_{S1} = Y_{s1} + \mu_{S1,g_{S1}^{opt}}(\mathbf{H}_{S1}) \times I(d_{S1} = 1)$ for the first step.

To reduce the accumulated bias from the conditional mean models, instead of using the model-based values under optimal future treatments $\widehat{\mu}_{sj,d_{sj}}(\mathbf{H}_{sj}) = \widehat{E}[PO_{sj} \mid D_{sj} = d_{sj}, \mathbf{H}_{sj}]$ from PO_{sj}, we use the actual observed intermediate outcomes plus the expected future loss (or gain) due to the sub-optimal treatments as the modified pseudo-outcome PO'_{sj} (*Huang et al.*, 2015). Specifically, the modified pseudo-outcome of the last stage is $PO'_{S2} = Y_{S2}$, $PO'_{S1} = Y_{S1} + \mu_{S2,g_{S2}^{opt}}(\mathbf{H}_{S2}) - \mu_{S2,D_{S2}}(\mathbf{H}_{S2}) + Y_{S2}$ and for any $s < S$, $j = 1$ or 2,

$$\begin{aligned} PO'_{sj} = \sum_{r=s+1}^{S} \Big[& \mu_{r1,g_{r1}^{opt}}(\mathbf{H}_{r1}) - \mu_{r1,D_{r1}}(\mathbf{H}_{r1}) + Y_{r1} \\ & + I[d_{r1} = 1][\mu_{r2,g_{r2}^{opt}}(\mathbf{H}_{r2}) - \mu_{r2,D_{r2}}(\mathbf{H}_{r2}) + Y_{r2}] \Big] \\ & + Y_{sj} + I[j = 1]I[d_{sj} = 1]\big[\mu_{s2,g_{s2}^{opt}}(\mathbf{H}_{s2}) - \mu_{s2,D_{s2}}(\mathbf{H}_{s2}) + Y_{s2}\big] \end{aligned} \tag{2.5}$$

In particular, if the subject undergoes the test at stage s, i.e., $d_{s1} = 1$, he/she might benefit from the potential subsequent treatment within that stage via the optimization of the future treatment step. If the subject does not receive the test at stage s, then his/her future optimized counterfactual outcome can only be optimized through the

optimal actions of the future stages.

We propose to implement SAT-Learning through a modified version of a tree-based reinforcement learning method (T-RL) (*Tao et al.*, 2018), which employs the classification and regression tree (CART) proposed by *Breiman et al.* (1984). In the nested DTR setting, we need to include the step-wise adjustment to account for the nested test-and-treat nature. Thus, we developed a modified tree-based algorithm to implement SAT-Learning for estimating the optimal nested DTR. Traditionally, the decision tree of CART is built to choose a split that would have the purest child nodes. The purest node means having the lowest misclassification rate among all possible nodes. Thus, purity is a crucial measure to grow a decision tree. Different from CART, SAT-Learning at each node selects the split to improve the counterfactual mean reward, which can serve as a measure of purity in nested DTR trees, and then maximizes the population's counterfactual mean reward of interest. Similarly as in T-RL, to estimate the optimal DTR, we use a purity measure for SAT-Learning based on the augmented inverse probability weighting (AIPW) estimator of the counterfactual mean outcome.

In the process of partitioning of this tree-based reinforcement learning method, for a given partition ω and ω^c of node Ω, let g_{sj,ω,d_1,d_2} denote the decision rule that assigns a single test/treatment action d_1 to all subjects in ω and treatment d_2 to subjects in ω^c at stage s step j $(1 \leq s \leq S, j = 1, 2)$. Then the purity measure can be defined as

$$\mathcal{P}_{sj}(\Omega, \omega) = \max_{d_1, d_2 \in \mathcal{D}_{sj}} \mathbb{P}_n \Big[\sum_{d_{sj}=1}^{K_{sj}} \widehat{\mu}^{AIPW}_{sj,d_{sj}}(\mathbf{H}_{sj}) I\{g_{sj,\omega,d_1,d_2}(\mathbf{H}_{sj}) = d_{sj}\} I(\mathbf{H}_{sj} \in \Omega) \Big], \tag{2.6}$$

where $\mathbb{P}_n$ is the empirical expectation operator and $\mathbb{P}_n\{\widehat{\mu}^{AIPW}_{sj,d_{sj}}(\mathbf{H_{sj}})\}$ is the AIPW

estimator of the counterfactual mean outcome with

$$\widehat{\mu}^{AIPW}_{sj,d_{sj}}(\mathbf{H}_{sj}) = \frac{I(D_{sj} = d_{sj})}{\widehat{\pi}_{sj,d_{sj}}(\mathbf{H_{sj}})} Y_{sj} + \left\{ 1 - \frac{I(D_{sj} = d_{sj})}{\widehat{\pi}_{sj,d_{sj}}(\mathbf{H_{sj}})} \right\} \mu_{sj,d_{sj}}(\mathbf{H_{sj}}). \tag{2.7}$$

In Eqn (2.7), the propensity score model is denoted as $\pi_{sj,d_{sj}}(\mathbf{H_{sj}})$ and the conditional mean model is denoted as $\mu_{sj,d_{sj}}(\mathbf{H_{sj}})$. Under the foregoing three causal inference assumptions, $\mathbb{P}_n\{\widehat{\mu}^{AIPW}_{sj,d_{sj}}(\mathbf{H_{sj}})\}$ is a consistent estimator of the counterfactual mean outcome $E\{Y^*(d_{sj})\}$ if either the propensity score model $\pi_{sj,d_{sj}}(\mathbf{H_{sj}})$ or the conditional mean model $\mu_{sj,d_{sj}}(\mathbf{H_{sj}})$ is correctly specified. Thus this AIPW estimator is doubly robust for estimating the counterfactual mean outcome of the population (*Tao and Wang*, 2017).

In our nested step-adjusted multi-stage setting, for the last step of the last stage, $S2$, we have Y_{S2} in Eqn (2.7) as the observed reward of the last step of the last stage. For other stage s step j before the last one ($1 \leq s < S, j = 1, 2$ or$s = S\,, j = 1$), Y_{sj} in Eqn (2.7) is replaced with PO'_{sj}, the corresponding pseudo-outcome defined in Eqn (2.5).

In the process of maximizing $\mathcal{P}_{sj}(\Omega, \omega)$, the possible split ω of a given node Ω should be either a subset of a categorical covariate categories or values that are not larger than the threshold. The best criteria $\widehat{\omega}^{opt}$ to split a given node is a partition that is able to maximize the improvement in the purity, $\mathcal{P}_{sj}(\Omega, \omega) - \mathcal{P}_{sj}(\Omega)$, where $\mathcal{P}_{sj}(\Omega)$ is for the situation where we assign the same single test/treatment action to all subject in Ω, i.e., no splitting. To control the overfitting and also make practical and meaningful splits, a positive integer n_0 is specified as the minimal node size and a positive constant λ is also provided as a threshold for the meaningful improvement. Besides the two given constant values λ and n_0, we apply similar *Stopping Rules* as in *Tao et al.* (2018) to grow and split the tree. Our Stopping Rules can be found in the *Appendix* as **Algorithm** 5. The depth of a node mentioned in the stopping

rules is defined as the number of edges from the node to the tree's root node, and a root node has a depth of 0. The nested SAT-Learning algorithm given the above purity measures and stopping rules of the partitioning is presented in **Algorithm** 1 with details. Note the essential difference between steps *j=2* and *j=1* is that different subjects are included into the calculation of the AIPW estimator. Only the subjects who have taken the test at stage s, i.e., $d_{s1} = 1$, contribute to the optimization of their subsequent treatments.

Algorithm 1 Implementation Steps of SAT-Learning

Stage s Start the algorithm with $s = S$ *Within Stage s:*

(1.1) Set $j = 2$ and only use the data with $d_{sj} = 1$

(1.2) Obtain AIPW estimates $\widehat{\mu}^{AIPW}_{sj,d_{sj}}(\mathbf{H}_{sj}), d_{sj} = 1, \ldots, K_{sj}$

(1.3) Set $m = 1$ at root node $\Omega_{sj,m}$

(1.4) At node $\Omega_{sj,m}$, evaluate the *Stopping Rules*. If stop, assign a single best treatment

$$arg \max_{d_{sj} \in \mathcal{D}_{sj}} \mathbb{P}_n[\widehat{\mu}^{AIPW}_{sj,d_{sj}}(\mathbf{H}_{sj}) I(\mathbf{H}_{sj} \in \Omega_{sj,m})]$$

to all subject in $\Omega_{sj,m}$. Otherwise, split $\Omega_{sj,m}$ into child nodes $\Omega_{sj,2m}$ and $\Omega_{sj,2m+1}$ by $\widehat{\omega}^{opt}$.

(1.5) Set $m = m + 1$ and repeat (1.4) until all nodes are terminal.

(2.1) Set $j = 1$ and use the full data and restrict the available nodes' values according to $\mathcal{P}_{s2}(\Omega, \omega)$

(2.2) Repeat Steps (1.2)-(1.5)

Next Stage: Set $s = s - 1$ and repeat ***Stage s:*** (1.1)-(2.2), stop if $s = 1$.

When implementing SAT-Learning process, the propensity score $\widehat{\pi}_{sj,d_{sj}}(\mathbf{H}_{sj})$ in Eqn (2.7) can be estimated by a multinomial logistic regression model. This working model could incorporate linear main effect terms from history $\mathbf{H}_{sj}$ and summary

variables or interaction terms based on prior scientific knowledge from individual history $\mathbf{H}_{sj}$. For continuous outcome, the conditional mean estimates $\widehat{\mu}_{sj,d_{sj}}(\mathbf{H}_{sj})$ in Eqn (2.7) could be obtained either from a linear parametric regression model or from other off-the-shelf non-parametric machine learning methods, such as random forests or support vector regression, depending on the history $\mathbf{H}_{sj}$ and the test/treatment action D_{sj}. For estimating the conditional mean model for binary or other count outcomes, one could use a generalized linear models or other generalized classification tools in machine learning.

2.5 Simulation Studies

2.5.1 Simulation Studies to Evaluate the General Test-and-treat Nested DTR

We generate simulation study data that mimic the real-world observational test-and-treat study. We assume a two-stage two-step nested dynamic treatment regime, using D_{sj} with subscript value $s = 1, 2$ to represent the stage and $j = 1, 2$ to represent the test and treatment action within each stage. More specifically, we set two options in the test step as $d_{s1} = 1$ or 0 to indicate receiving the test or not, and three treatment options in the treatment step as $d_{s2} = 0, 1$ or 2. We further define the outcome of interest as the sum of intermediate rewards from each stage and step, i.e., $Y = Y_{11} + Y_{12} + Y_{21} + Y_{22}$. The underlying optimal treatment is supposed to have the largest expected reward. The other two sub-optimal treatments have lower expected rewards. We further consider two cases. One is that the expected reward from the two sub-optimal treatments are equal while in the other case, the expected reward of the two sub-optimal treatments are different. Therefore, in the second case, the sub-optimal reward losses are different because patients may lose more treatment benefit due to choosing one sub-optimal treatment compared to another.

When the test step initiating each intervention step is not expensive or invasive, more patients tend to choose such a test because they might benefit from knowing the test result for the long term disease control purpose. However, when the lab test is unpleasant and costly, such as a prostate biopsy test, the patients would hesitate to take it. Therefore, when generating data we consider three scenarios based on the patients' willingness to receive the exam by modifying the parameters to set the ratio of having or not having the test as 1:1, 2:1, and 1:2, which correspond to the equal preference, more likely and less likely to take the exam, respectively. For these three scenarios, three covariates, X_1 to X_3, generated as the baseline covariates follow $N(0,1)$. Two correlated covariates, X_4 and X_5, are generated as time-varying biomarkers which are measured just before the decision time of the test step within each stage. $(X_4, X_5)^{'} \sim N(\mu, \Sigma)$, where $\mu = (0,0)^{'}$ and $\Sigma = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$. After the test step of each stage, the covariates X_{12} and X_{22} mimic the test results that contribute to the treatment decision nested within each test decision with other covariates. Typically, the test results, such as biopsy results, are of great importance to the treatment decision making. X_{12} and X_{22} follow the distribution of $N(0,1)$. Details of parameter setting are as follows:

Stage 1: The test decision variables, $D_{11} \sim Bernoulli(\pi_{11,1})$ with $\pi_{11,1} = \exp(0.6X_3 - 0.2X_2 + X_4)/(1 + \exp(0.6X_3 - 0.2X_2 + X_4))$. The reward of step 1 of stage 1 is generated as $Y_{11} = X_4^2 + (0.5X_3 + 3)^2 \times I[g_{11}^{opt}(\mathbf{H}_{11}) = D_{11}] - 3|X_1|I(D_{11} = 1) + \epsilon_{11}$ with optimal regimes defined as

$$g_{11}^{opt}(\mathbf{H}_{11}) = \begin{cases} I(X_1 > -0.5)I(X_4 \leq 0.3) & \text{for Scenario 1} \\ I(X_1 > -0.8)I(X_4 \leq 1) & \text{for Scenario 2} \\ I(X_1 > 0.3)I(X_4 \leq 1.3) & \text{for Scenario 3,} \end{cases}$$

and $\epsilon_{11} \sim N(0,1)$. The Scenarios 1, 2, and 3 corresponds to patients' equal preference,

more likely, and less likely to take the test, respectively. For patients who have taken the test, i.e., $D_{11} = 1$, we further generate the treatment assignment D_{12} for them as $D_{12} \sim Multinomial(\pi_{12,0}, \pi_{12,1}, \pi_{12,2})$ with $\pi_{12,0} = 1/(1 + \exp(0.5X_{12} - 0.2X_2) + \exp(0.2X_4 + 0.3X_3))$, $\pi_{12,1} = \exp(0.5X_{12} - 0.2X_2)/(1 + \exp(0.5X_{12} - 0.2X_2) + \exp(0.2X_4 + 0.3X_3))$ and $\pi_{12,2} = \exp(0.2X_4 + 0.3X_3)/(1 + \exp(0.5X_{12} - 0.2X_2) + \exp(0.2X_4 + 0.3X_3))$. Also, $Y_{12} = I[D_{12} = g_{12}^{opt}(\mathbf{H}_{12})](2X_{12} + 3X_2)^2 + (X_1 + X_3 * 2 + X_4) + Y_{11}/3 + \epsilon_{12}$ for equal sub-optimal reward loss; and

$$\begin{aligned} Y_{12} =& I[D_{12} = g_{12}^{opt}(\mathbf{H}_{12})](2X_{12} + 3X_2)^2 + (X_1 + X_3 * 2 + X_4) + Y_{11}/3 \\ &+ 0.5I(D_{12} = 1)[I(g_{12}^{opt}(\mathbf{H}_{12}) = 1) - 1] + 1.2I(D_{12} = 2)[I(g_{12}^{opt}(\mathbf{H}_{12}) = 2) - 1] \\ &+ \epsilon_{12} \end{aligned}$$

for unequal sub-optimal reward loss with $\epsilon_{12} \sim N(0, 1)$. The tree-type optimal regime at step 2 is specified as

$$g_{12}^{opt}(\mathbf{H}_{12}) = \begin{cases} 0 & X_{12} > 0.2 \\ 1 & X_1 > -0.7, X_{12} \leq 0.2 \\ 2 & \text{otherwise.} \end{cases}$$

Stage 2: We generate the test decision of stage 2, $D_{21} \sim Bernoulli\,(\pi_{21,1})$ with $\pi_{21,1} = \exp(0.5X_1 - 0.6X_2 + X_3)/(1 + \exp(0.5X_1 - 0.6X_2 + X_3))$. The reward of stage 2 step 1 is generated as $Y_{21} = X_5^2 + 2X_1 + (X_3 + 3.2)^2 I[g_{21}^{opt}(\mathbf{H}_{21}) = D_{21}] - 3I(D_{21} = 1) + \epsilon_{21}$ with $\epsilon_{21} \sim N(0, 1)$. The optimal regime $g_{21}^{opt}(\mathbf{H}_{21})$ is specified as

$$g_{21}^{opt}(\mathbf{H}_{21}) = \begin{cases} I(X_1 \leq -0.3) + I(X_1 > -0.3)I(X_5 \geq 1) & \text{for Scenario 1} \\ I(X_1 \leq 0.4) + I(X_1 > 0.4)I(X_5 \geq 1.2) & \text{for Scenario 2} \\ I(X_1 \leq -0.8) + I(X_1 > -0.8)I(X_5 \geq 1) & \text{for Scenario 3.} \end{cases}$$

Among the patients who have had the test, i.e., $D_{21} = 1$ we generate their treatment assignment D_{22} for the second step of stage 2. Specifically, we generate treatment $D_{22} \sim Multinomial(\pi_{22,0}, \pi_{22,1}, \pi_{22,2})$ with $\pi_{22,0} = 1/(1 + \exp(0.35X_{22} - X_5) + \exp(0.3X_2 + 0.2X_3))$, $\pi_{22,1} = \exp(0.35X_{22} - X_5)/(1 + \exp(0.35X_{22} - X_5) + \exp(0.3X_2 + 0.2X_3))$, and $\pi_{22,2} = \exp(0.3X_2 + 0.2X_3)/(1 + \exp(0.35X_{22} - X_5) + \exp(0.3X_2 + 0.2X_3))$. The reward of stage 2 step 2 is generated as $Y_{22} = 3I[D_{22} = g_{22}^{opt}(\mathbf{H}_{22})] + Y_{21} + (2 + X_4X_5 + X_3) + \epsilon_{22}$ for equal sub-optimal reward loss; and

$$\begin{aligned} Y_{22} =& (3 + X_{22})I[D_{22} = g_{22}^{opt}(\mathbf{H}_{22})] + Y_{21} + (2 + X_4X_5 + X_3) \\ &+ 2I(D_{22} = 1)[I(g_{22}^{opt}(\mathbf{H}_{22}) = 1) - 1] + I(D_{22} = 2)[I(g_{22}^{opt}(\mathbf{H}_{22}) = 2) - 1] + \epsilon_{22} \end{aligned}$$

for unequal sub-optimal reward loss, and $\epsilon_{22} \sim N(0, 1)$. The optimal treatment regime for stage 2 $g_{22}^{opt}(\mathbf{H}_{22})$ is specified as

$$g_{22}^{opt}(\mathbf{H}_{22}) = \begin{cases} 0 & X_{22} > 0.5 \\ 1 & X_{22} \leq 0.5, X_5 < 0.3 \\ 2 & \text{otherwise.} \end{cases}$$

Table 2.1 summarizes the simulation study results across different scenarios as described above. Our SAT-Learning method for estimating the optimal DTR involves a doubly robust semi-parametric estimator, therefore our simulations also try to demonstrate such robustness. In addition to having one estimation scheme with the conditional mean model and the propensity score model both correctly specified, we consider two more schemes with either the propensity score model or the conditional mean model mis-specified by omitting some of the covariates of the true form. We consider a sample size of either 1000 or 2000 for the training dataset, and a sample size of 2000 for the validation, and repeat the simulation 500 times. The training dataset is used to estimate the optimal regime and then predict the optimal test-and-

Table 2.1: Simulation results for the general test-and-treat case for the equal and unequal reward loss for sub-optimal treatment options: two intervention stages, three treatment options at each stage nested within the exam at each stage with 500 replications, and n=1000 or 2000.

Sample Size	Sub-optiomal Reward		Scenario 1 (1:1) *opt*%	Scenario 2 (2:1) *opt*%	Scenario 3 (1:2) *opt*%
N=1000	Equal Loss	(a)	90.1(7.4)	86.1(9.2)	91.9(6.3)
		(b)	84.7(7.5)	81.0(7.9)	86.9(6.4)
		(c)	90.1(7.6)	86.3(9.3)	92.1(6.4)
	Unequal Loss	(a)	96.2(3.8)	96.3(4.0)	97.7(2.0)
		(b)	92.0(6.1)	87.5(12.5)	94.7(4.1)
		(c)	96.0(4.1)	96.2(4.4)	97.6(2.2)
N=2000	Equal Loss	(a)	91.2(7.5)	86.8(9.3)	93.2(6.4)
		(b)	85.8(6.6)	81.9(6.3)	88.2(6.1)
		(c)	91.1(7.5)	86.9(9.3)	93.2(6.4)
	Unequal Loss	(a)	96.9(3.4)	97.7(2.7)	98.2(1.8)
		(b)	96.6(3.6)	93.8(8.8)	97.7(1.7)
		(c)	96.9(3.4)	97.7(2.6)	98.1(1.8)

a. *opt*% show the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their underlying true optimal treatments
b. The optimal regimes are estimated by the proposed method when (a) the conditional mean model and the propensity score model are both correctly specified, (b) the conditional mean model is mis-specified and the propensity score model is correctly specified, and (c) the conditional mean model is correctly specified and the propensity score model is mis-specified.
c. Scenarios 1,2 and 3, correspond to the cases when the true ratios of preference for having the exam v.s. not having the exam among all patients are 1:1, 2:1 and 1:2.

treat decision in the validation dataset, where the underlying true optimal regimes are already known. The percentages of subjects correctly classified to the optimal test-and-treatment decision in both stages combined is denoted as *opt%*. The average *opt%* and the empirical standard deviation (SD) among the repetitions evaluate the performance.

In Table 2.1, the results of equal sub-optimal reward case demonstrate the loss due to sub-optimal equally inferior compared to the optimal choice. In scenario 1, where the subjects have an even preference of having test, under the sample size n=1000, 90.1% of the patients are correctly assigned to their optimal DTRs for both stages when both the conditional mean model and the propensity score model are correctly specified. When either the conditional mean model or the propensity score model is mis-specified, but not both, the overall performances are slightly worse,

but still reasonably satisfactory. More specifically, when either the propensity score model or the conditional mean model is incorrectly specified, we still get 90.1% and have 84.7% respectively. Similar trends are found in Scenario 2 and Scenario 3, and results improve as sample size goes to N=2000. However, when the reward loss is unequal among sub-optimal treatment options, the optimal regimes stand out among the candidate treatments more obviously according to our data generating process, therefore, it is easier for our proposed method SAT-Learning to distinguish the optimal treatment from sub-optimal ones. Thus, the simulation performance with varying sub-optimal loss is better than when the sub-optimal loss is equal, as expected.

2.5.2 A Special Case when the Treated Patients no Longer Need Further Test or Treatment

We conduct another simulation study for a special case when the treated patients no longer need test and treatment again. This simulation better mimics the monitoring and management in active surveillance for prostate cancer. Because of the significant side-effects of curative intervention and the asymptomatic nature of prostate cancer, according to the American Society of Clinical Oncology, patients with low-risk prostate cancer can consider active surveillance (*Tosoian et al.*, 2011; *Klotz et al.*, 2014). Active surveillance involves monitoring prostate cancer by regular exam in its localized stage until further treatment is needed to halt the disease at a curable stage. More specifically, the patients who have taken the biopsy test, only a small proportion of them would switch from the active surveillance to curative intervention. In the active surveillance, the patients who have been treated should be removed from the active surveillance cohort, because physicians consider that they no longer need to be treated and additional treatment is not provided and they are not eligible for the active surveillance. Therefore they should not be considered to

evaluate the subsequent test or treatment decision. We generate data under a two-stage nested DTR with two treatment options at each stage. We also modify the parameters in the data generating models to make the rates of taking the curative treatment equal to 5%, 15%, 20% and 25% in both stage. The higher the rate is, the more patients take the treatment and thus more patients will be removed from the surveillance afterwards. The detailed information of data generation can be found in the *Appendix*

The simulation results are summarized in Table 2.2. As the results show, because of the nice doubly robust property, the percentages of subjects correctly classified to their underlying truth both yield satisfying results even when either the propensity score model or the conditional mean model is mis-specified, but not both. Considering sample size N=1000 as an example, when 5% of tested patients have the curative treatment and then are removed from the active surveillance, 92.8% of them are correctly assigned to their optimal DTR for both stages when both the conditional mean model and the propensity score model are correctly specified. We also have 91.8% and 91.6% of the patients correctly classified to their optimal DTR when the propensity score model or the conditional mean model is misspecified respectively. As the treatment rate increases, we are able to estimate better optimal treatment rules from larger heterogeneous samples with more information. Therefore, it is easier for our proposed SAT-Learning to estimate the optimal regime from this more informative sample. Thus, the simulation performance with a higher treatment rate is slightly better than that for the lower rate case.

2.6 Application to Prostate Cancer Active Surveillance Data

We illustrate SAT-Learning using the prostate cancer Active Surveillance dataset from Johns Hopkins University (*Tosoian et al.*, 2011; *Inoue et al.*, 2018; *Lange et al.*, 2018). In this active surveillance study, enrollment of men with low risk prostate

Table 2.2: Simulation to mimic the monitoring and management of prostate cancer: two intervention stages, two treatment options at each stage nested within the exam at each stage with 500 replications, and n=1000 or 2000.

Treatment Rate		5% *opt*%	15% *opt*%	20% *opt*%	25% *opt*%
N=1000	(a)	92.8(4.9)	93.8(4.8)	94.9(4.5)	95.3(4.6)
	(b)	91.6(2.1)	93.2(1.9)	93.9(1.5)	94.3(1.4)
	(c)	91.8(5.6)	93.2(5.8)	94.0(5.9)	94.6(5.5)
N=2000	(a)	93.7(4.7)	94.9(4.6)	95.7(4.4)	96.7(3.7)
	(b)	92.6(1.9)	94.0(1.4)	94.5(1.2)	94.7(1.1)
	(c)	92.2(6.5)	93.9(6.6)	94.6(6.5)	95.5(6.1)

a. *opt*% show the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their underlying true optimal treatments

b. The optimal regimes are estimated by the proposed method when (a) the conditional mean model and the propensity score model are both correctly specified, (b) the conditional mean model is mis-specified and the propensity score model is correctly specified, and (c) the conditional mean model is correctly specified and the propensity score model is mis-specified.

c. Different treatment rates correspond to different proportions of patients who switch from active surveillance to curative treatment among those who have taken the biopsy test.

cancer started in 1995 and ended in 2015. Eligible subjects need to have PSA density less than 0.15 μg/L per mL, clinical stage T1c disease or lower, the Gleason score between 2 and 6, at most 2 positive biopsy cores, and at most 50% tumor in any single core, all of which made them low risk. The Johns Hopkins active surveillance protocol includes semiannual PSA and annual prostate biopsy. In the protocol, the primary reason that patients would be recommended to undergo definitive curative radiation therapy or surgery is if the biopsy result showed an adverse change compared to previous biopsies.

There is sufficient evidence that the approach of active surveillance, i.e. delaying curative treatment, for low-risk patients is safe (*Denton et al.*, 2019). The issue we will be considering is how it should be implemented. That is, rather than having an annual biopsy, as in the protocol, should it be more individualized, with the decision of whether to undergo a biopsy based on the available data at that time for that patient.

Not all the patients in the study followed the protocol. In the dataset we analyzed,

22% of patients did not have the scheduled biopsy of the first year and 5% of them did not have the biopsy in the first two years. Similarly for curative therapy, quite a number of patients did not follow the protocol. Such heterogeneity in the observed data allows us to apply the nested DTR method via our proposed SAT-Learning to decide at each stage whether the patient should have the biopsy, and if so, whether the treatment should be recommended based on the patients' individualized characteristics. In the analysis presented below we restrict the observational period to be from the diagnosis to year 4 and we make a two-year time unit for each stage, making two stages, stage 1 being from diagnosis to year 2 and stage 2 being from year 2 to year 4. We use D with subscript value $s = 1, 2$ to denote the decisions of the two stages, and $j = 1, 2$ to denote the biopsy and treatment actions within the stage. Thus, if the subject had a biopsy at the first stage, we denote $D_{11} = 1$, otherwise, $D_{11} = 0$. For those with biopsy i.e., $D_{11} = 1$, the treatment choice is recorded as D_{12}, 1 for treated and 0 for no treatment, and similarly for D_{21} and D_{22}. We note that once the patient is treated, no further biopsy or treatment will be observed. After the data preprocessing, 863 patients are kept in the dataset for the analysis, and of these 230 did receive curative treatment. More information regarding data preprocessing can be found in *Appendix.*

Although patients, in reality, are subject to different categories of treatments, such as prostatectomy, radiation therapy or hormone therapy, in this analysis, we combine all different kinds of treatments into one category (treated) to preserve a sufficient sample size for the treated subjects. Other patient characteristics, including age, race, baseline biopsy results, and baseline PSA were collected at the enrollment. As the active surveillance proceeded, the corresponding PSA changes and the follow-up biopsy results were also collected. In particular, the quantity of cancer, as measured by biopsy results, is based on both the number of needle cores containing cancer and the characteristic of the cancer tissue found within each single core (Gleason score).

How the individualized data was formatted to match the two year time interval for each stage is described in the *Appendix*. The reward outcome of interest was chosen to reflect long term disease status, and is defined as the proportion of PSA values which are less than 5 out of all the PSA observations collected from the end of year 4 after diagnosis to the end of study. This reward ranges from 0 to 1, with the lower values implying more undesirable risk of prostate cancer progression. This reward outcome only considers the disease prognosis based on PSA, and ignores the potential side effects brought by frequent biopsy and unnecessary intervention. Thus, we include penalties to discount the patient's reward to take into account possible side effects. More specifically, if the patient had a biopsy in either one of the two stages, his reward is reduced by a factor of 87% compared to the original reward. For the patient who has ever had treatment, the reward is reduced by a factor of 80% compared to his original reward.

To apply the proposed SAT-Learning algorithm to the active surveillance data described above, we use random forests for the conditional mean model and a logistic regression model for the propensity score model of every step within each stage. The estimated optimal test and treatment DTR of the two stages are shown in Figure 2.2 According to the estimated optimal DTR, at the first stage, men older than 56 are recommended for a biopsy test. Among those who are younger than 56 years old, the patients with most recent PSA higher than 3.6 are also recommended for a biopsy test. Among those doing the biopsy test, patients with the most recent PSA higher than 3.1 and having biopsy test showing any cancer are recommended for the treatment. At the second stage, the men whose PSA change from beginning of year 2 is larger than 1.3 are recommended for the biopsy test. For those who take the biopsy, if their most recent PSA is higher than 3.2 or the biopsy result has more than one biopsy core needle showing cancer positive, we recommended the physician to offer them the treatment. The standard practice in deciding on curative treatment depends

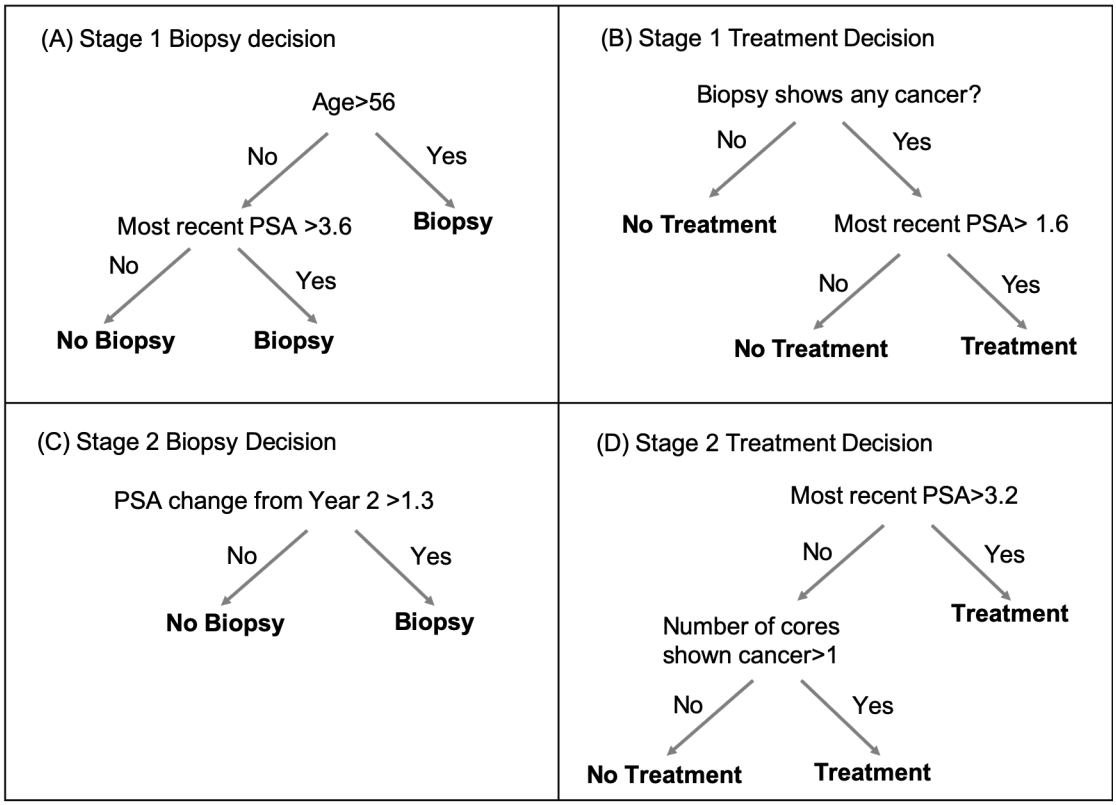

Figure 2.2: Estimated optimal DTR for JHU prostate cancer active surveillance data via SAT-Learning algorithm. The trees show how to provide optimal regime at every step based on the individualized characteristics for (A) stage one biopsy decision, (B) stage one treatment decision if biopsy was taken in stage one,(C) stage two biopsy decision and (D) stage two treatment decision if the biopsy was taken in stage two.

primarily on whether the Gleason grade on the biopsy is greater than or equal to 7. In contrast, the DTR we estimated involves more variables and changes from one stage to the next, so is more individualized. It is also notable that the Gleason thresholds in the above DTR are lower than in standard practice, which is consistent with a suggestion in the literature (*Moyer*, 2012). The reward we use, long run low PSA values, certainly does influence the estimated DTR, which involves lots of decisions based on the current PSA values. The estimated tree-based DTR presented in Figure 2.2 is also sensitive to the discount factor 87% and 80% which are used to penalize the reward. Other rewards would have given different DTRs. The reward we use of long-run PSA values can be considered as a proxy for clinical meaningful "good" outcome. An ideal reward would have involved long term good quality of life and absence of prostate cancer recurrence. But data to construct such a reward is not available for this study. A sensitivity analysis with a modified reward is presented in the *Appendix*.

2.7 Discussion

Motivated by the embedded nature of the diagnosis and treatment procedures, we have developed a nested DTR framework, with the treatment decision nested within the test decision in a multi-stage setting, and implemented the estimation of the optimal nested DTR using a step-adjusted tree-based reinforcement learning method (SAT-Learning). This nested DTR framework considers the test decision and the nested treatment decision in the same stage and develops the optimal nested DTRs to maximize the expected long-term rewards, such as disease control. This kind of test-and-treat strategy has been considered previously in the health policy literature (*Trikalinos et al.*, 2009). These methods discussed the importance of the problem, and the need to accumulate data. They also suggested solutions that focused on the population level, but not in a rigorous mathematical framework. Our proposed

method follows the framework of DTR, which enables physicians to repeatedly tailor test and treatment decisions based on each individual's time-varying health histories, and thus provides an effective tool for the personalized management of disease over time.

SAT-Learning, our proposed method to solve the nested step-adjusted DTR problem, can potentially be implemented via modifying other learning methods that have been considered in DTR literature. However, by using a modified T-RL algorithm (*Tao et al.*, 2018), SAT-Learning is more straightforward to implement, understand and interpret, and capable of handling various data without distributional assumptions. Additionally, the doubly robust AIPW estimator that we utilize in the purity measure in the tree structure also helps improve the robustness of our method against model mis-specifications.

Several developments and extensions can be explored in future studies. One possible exploration lies on dealing with potentially contradictory multiple outcomes. In SAT-Learning, we consider a nested step-adjusted DTR to reduce the pain and potential infections from frequent biopsy tests, but maintain an effective and in-time treatment to control disease progression. If efficacy is the only purpose, one would expect more frequent tests and more aggressive treatment regardless of possible side effect, but in the meantime, patients might experience more side effects. The desire for efficacy and the desire for less side effects in fact contradict each other. In clinical practice, physicians are often interested in balancing multiple competing clinical outcomes, such as overall survival, patient preference, quality of life and financial burden (*Butler*, 2016). In order to balance these multiple potentially contradictory objectives, we applied a different discount factor to the patient rewards for different side effects in the application to the JHU Prostate Cancer Active Surveillance data. Other statistical methods have been developed to trade-off between multiple contradictory outcomes (*Laber et al.*, 2014; *Lizotte and Laber*, 2016). One can further incorporate

these multiple objective optimization functions into our framework of nested DTR for future research. Another possible exploration may be considering all available actions when the preference of multiple outcomes varies (*Lizotte et al.*, 2012), which would give more comprehensive information about how the optimality of an action would be changed if the preference is modified. Sensitivity analyses can be done on the optimal regimes and would provide further guidance for the decision maker on developing a more flexible regime among all the available intervention strategy choices.

2.8 Acknowledgement

The authors are grateful to Dr. Brian Denton for providing the prostate cancer dataset. This research was partially supported by National Institutes of Health grant CA199338.

CHAPTER III

Kernel-Involved-Dosage-Decision Learning Method for Estimating the Optimal Dynamic Treatment Regimes

3.1 Introduction

Dose-finding has a critical role in clinical research. The optimal dose for a drug should be balanced between the safety and efficacy requirements (*Schmidt*, 1988). There has been a great deal of literature on the dose-finding methods for clinical studies. (*Braun et al.*, 2016; *Cheung*, 2011; *Thall and Russell*, 1998; *Thall and Cook*, 2004; *Thall et al.*, 2007; *Thall*, 2008). Moreover, as the interest in precision medicine increases (*Collins and Varmus*, 2015), scientists realize that a one-size-fits-all optimal dose recommendation is not appropriate for treating heterogeneous patients. Recent dose-finding methods have also evolved to account for the patients' characteristics and estimate the personalized optimal dose for each individual (*Li et al.*, 2019; *Guo and Yuan*, 2017; *Thall et al.*, 2008; *Xu et al.*, 2018; *Rich et al.*, 2016).

In addition to a more personalized dose assignment, treating patients with chronic disease often includes more than one cycle of treatments. Physicians may treat patients routinely in every stage based on the current status of the time-varying biomarkers and other health characteristics. Typically, a sequence of decision rules should

be considered as a whole treatment regime, rather than several independent decision rules. This kind of sequence of treatments is not only a strategy for the management of chronic disease, but also a device for achieving better efficacy while controlling the toxicity (*El Naqa et al.*, 2018). For example, patients with liver cancer were treated at Michigan Medicine with adaptive stereotactic body radiation therapy (SBRT) with an intra-treatment evaluation. Instead the total radiation dose at the beginning of radiation therapy,only 60% of the total planned dose is given during the first month, with the remaining 40% dose given or partially given after a one-month break, during which the toxicity and the patient's tolerability were carefully evaluated (*Feng et al.*, 2013). This novel adaptive treatment paradigm showed its benefit in achieving lower toxicity while maintaining a comparable local tumor progression compared to the non-adaptive radiation therapy (*Feng et al.*, 2018). In this adaptive SBRT study, there is not only one decision, but a sequence of two decision rules, one per stage, where the second one is determined by considering the observed updated medical characteristics after the first treatment dose. Such a paradigm is known as a dynamic treatment regime (DTR) (*Murphy et al.*, 2001; *Murphy*, 2003; *Wang et al.*, 2012), which consists of decision rules, one per stage, mapping individualized patient characteristics to a dose. However, most multiple-stage dose-finding clinical studies, while using the individualized adaptive treatment paradigm, only evaluate the dose response as if the patient's outcome was due to a certain stage alone, rather than the entire DTR (*Lee et al.*, 2015). In addition, these studies do not take full advantage of the intra-treatment information between treatment stages (*El Naqa et al.*, 2018). Overlooking this dynamic treatment regime structure may result in identifying a suboptimal dosage regime. Therefore, there is a need for a new statistical learning method that is capable of estimating the optimal dynamic dosage regime within the setting of a DTR. In the remaining part of the article, we will refer to dynamic treatment regimes with continuous dose treatments as dynamic dosage regime.

Extensive statistics literature exists on estimating the optimal DTR under the setting of multiple-stage multiple-treatment (*Zhang et al.*, 2015; *Laber and Zhao*, 2015; *Tao and Wang*, 2017; *Schulte et al.*, 2014; *Zhang et al.*, 2013). However, extending the current methods to estimate an optimal dynamic continuous dosage regime is not trivial. Using the observed outcomes from the patients whose dose assignment follows a specific rule is not applicable for continuous doses, because there could be an infinite number of treatment options for a given dose interval. Specifically, unlike the multiple-treatment DTR problem, only a few patients may be observed using the given dose level because that the dose level follows a continuous distribution – of which the probability of observing a certain rule-assigned dose is zero. Thus, inadequate methods have been developed for estimating the optimal continuous dosage regime in a multiple-stage setting.

Analogous to the DTR problem in a multiple-treatment multiple-stage setting, *Lee et al.* (2015) extended the Q-Learning (*Watkins and Dayan*, 1992), a commonly used reinforcement learning method, to estimate the dynamic treatment regime for continuous treatment. Q-Learning involves a two-step regression-based approach with a regression model fitted in the first step. In the second step, by maximizing the expected mean outcome from the first step, the optimal treatment for a given history can be predicted. However, Q-learning is susceptible to potential over-fitting of the first step regression model, which may lead to a sub-optimal DTR (*Lee et al.*, 2015). Moreover, several modifications to outcome weighted learning (OWL) (*Zhao et al.*, 2012), a well-known direct method in the setting of finite treatment-option DTR, have also been developed to accommodate the estimation of the optimal dosage regime. To deal with the individualized treatment rule problem in an ordinal treatment setting, *Chen et al.* (2018) proposed the Generalized Outcome Weighted Learning (GOWL), which could potentially be applied for estimating the optimal dosage regime. However, the dose options dealt by GOWL remain restricted within a finite number of

options while most dose treatment options observed in practice are on a continuum. *Chen et al.* (2016) also modified the OWL for estimating a personalized optimal dose in a continuous scale (denoted as CZK hereafter). However, analogous to its originated method, OWL, CZK is susceptible to attempting to retain the actual observed dose, because only an observation in which the observed dose is close to the estimated optimal dose can contribute to the loss function. In addition, the estimated individualized dosage decision is affected by a simple shift of the outcome. Moreover, the CZK and GOWL are flexible in their forms but difficult to interpret since the dosage decisions are derived from "black boxes." Therefore, it is more desirable to have interpretable dosage regimes for physicians to understand and apply. *Laber and Zhao* (2015) also developed a method to solve this dosage strategy problem by using a tree-based method (denoted as LZ hereafter). LZ is capable of providing interpretable tree-based decision rules. However, it relies on the correct specification of the outcome regression model, and therefore LZ is fragile due to model mis-specifications.

To overcome the limitations of the existing methods, we propose a robust and interpretable personalized dose-finding method, kernel-involved-dosage-decision learning (KIDD-Learning). At each stage, KIDD-Learning combines a non-parametric estimation of the dose-response function with an interpretable tree-based decision rule, to estimate the optimal dynamic dosage regimes in a multiple-stage setting, using observational data. The whole dynamic dosage regime is estimated by backwards inductions.

The remaining parts of this paper are organized as follows: In section 3.2, we formalize the problem of identifying the optimal dynamic dosage regime using the counterfactual causal inference framework in a multiple-stage setting. Section 3.3 develops the KIDD-Learning method to solve the dose-finding problem and describes the detailed implementation of KIDD-Learning. Section 3.4 extends the methods to handle a time-to-event outcome to accommodate the censored data we used in

the application section. The simulation studies and an application to the Michigan Medicine SBRT liver cancer patient dataset are provided in section 3.5 and 3.6. A discussion concludes this method in section 3.7.

3.2 Data and Formulation of Dynamic Dosage Regime

3.2.1 Statistical Problem for Optimizing Dynamic Dosage Regime

To address the dynamic continuous dosage regime problem above, we consider a multiple-stage continuous dosage decision framework with T decision stages. At each stage t, let $D_t \in \mathcal{D}_t$ denote the continuous dose value of the treatment taken at stage t with observed value d_t. Without loss of generality, we further assume $\mathcal{D}_s = [0, 1]$, and $d_t \in \mathcal{D}_t$.

The patient's accumulated history between stage $t-1$ and t is denoted as X_t. We use the over bar with subscripts t to denote a vector of a variables' history up until stage t, i.e., $\overline{\mathbf{X}}_t = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t)$. Similarly, the treatment history until stage t can be denoted as $\overline{\mathbf{D}}_t = (D_1, \ldots, D_{t-1})$. We use Y_t to denote the intermediate outcome at the end of stage t, and thus the overall outcome vector is $(Y_1, Y_2, \ldots, Y_T)$ and the outcome history $\overline{\mathbf{Y}}_{t-1} = (Y_1, \ldots, Y_{t-1})$. The outcome of interest Y is a function of all intermediate outcomes, i.e., $Y = f(Y_1, Y_2, \ldots, Y_T)$, where $f(\cdot)$ is a prespecified function (e.g., sum). Y is assumed to be bounded and higher value of Y is more desirable. The history $\mathbf{H}_t$ is defined as the treatment and medical history prior to treatment decision D_t. Specifically, we denote $\mathbf{H}_t = \{(\overline{\mathbf{D}}_t, \overline{\mathbf{X}}_t, \overline{\mathbf{Y}}_{t-1})\}_{i=1}^n$. The observed data consist n i.i.d. trajectories are therefore denoted as $\{D_{t,i}, \mathbf{H}_{t,i}, Y_{t,i}\}_{t=1}^T$ across patient index i. For brevity, we suppress the subject index i in the following text when no confusion exists. The goal of our method is to use the observed data to find the optimal dosage regime that determines what dose a patient should received at each stage based on his/her medical history. Formally, the dynamic dosage regime for the contin-

uous treatment is defined as a personalized decision rule sequence $\mathbf{g} = (g_1, g_2, \ldots, g_T)$ that maps the observed history $\mathbf{H}_t$ about the patient's health characteristics to the dose assignment at stage t, i.e.,

$$\mathbf{H}_t \mapsto g_t(\mathbf{H}_t) \in \mathcal{D}_t,\ 1 \leq t \leq T$$

To identify the optimal dynamic dosage regime from the observed data, we follow the counterfactual framework of causal inference proposed by *Robins* (1986). Let $Y^*(\mathbf{g})$ be the counterfactual outcome if a patient follows the dynamic treatment regime $\mathbf{g}$. $E\{Y^*(\mathbf{g})\}$ is the expectation of the counterfactual outcome with respect to the distribution of patients' history if the entire population had follow the dosage regime $\mathbf{g}$. The optimal $\mathbf{g}^{opt}$ is a sequence of decision rules that leads to the optimal value of $E\{Y^*(\mathbf{g})\}$, such that,

$$E\{Y^*(\mathbf{g}^{opt})\} \geq E\{Y^*(\mathbf{g})\},\ \forall\, \mathbf{g} \in \mathcal{G}, \tag{3.1}$$

where $\mathcal{G}$ is the set of all potential dosage regimes under consideration.

3.2.2 Using the Observational Data to Estimate the Optimal Dosage Regime $\mathbf{g}^{opt}$

According to the problem described in Eqn (3.1), the expectation of counterfactual is used to estimate the optimal dose. However, the counterfactual outcomes of subjects are not available in most observational studies. Therefore, we have to apply the framework of causal inference to connect the counterfactual outcome we desire with the observational study data we observe. Specifically, the method to find the whole optimal dynamic dosage regime includes backward induction; therefore, we start the mathematical formulation from the final stage T in a reverse sequential order to the previous treatment stages.

At the final stage, let $Y^*(D_1, D_2, \ldots, D_{T-1}, d_T)$, or $Y^*(d_T)$ for brevity, be the counterfactual outcome if a patient makes decision d_T conditional on his/her previous histories. To connect the counterfactual outcome with the observed data $\{D_T, \mathbf{H}_T, Y\}$, we make the following standard causal inference assumptions:

Assumption 1 *Consistency*: $D_T = d_T$ implies $Y = Y^*(d_T)$. This assumption ensures the observed outcome is the same as the counterfactual outcome under the treatment actually assigned;

Assumption 2 *Exchangeability*: $Y^*(d_T) \perp D_T \mid \mathbf{H}_T, \forall\, d_T \in \mathcal{D}_T$, where $\perp$ denotes statistical independence. Conditional on the previous history $\mathbf{H}_T$, the counterfactual outcome under a certain dose d_T is independent of the dose choice;

Assumption 3 *Positivity*: $\pi(d_T \mid \mathbf{H}_T) \geq \pi_{min} > 0$ for all possible $\mathbf{H}_T$, where $\pi(d_T \mid \mathbf{H}_T) = \frac{\partial P(D_T \leq d_T \mid \mathbf{H}_T)}{\partial d_T}$ is the conditional treatment density given history $\mathbf{H}_T$ and π_{min} is a given positive number. It ensures that every patient has some chance of receiving a certain treatment d_T.

Under the assumptions above, we are able to identify the optimal regime using the observed data. At the final stage T, the expectation of counterfactual outcome $Y^*(d_T)$ can be identified with observed data as

$$E\{Y^*(d_T)\} = E\{E(Y \mid D_T = d_T, \mathbf{H}_T)\} = \int_{\mathcal{H}_T} E(Y \mid D_T = d_T, \mathbf{H}_T = \mathbf{h}_T)\, \mathrm{d}P(\mathbf{h}_T).$$

When studying the continuous exposure, especially using observational study dataset, dose-response function is a natural way visualizing the dose-response relationship and describing the dosage effect, rather than using a scalar (*Kennedy et al.*, 2017). Therefore, we denote $\theta_T(d_T) = E\{Y^*(d_T)\}$, and $\theta_T(\cdot)$ be the dose-response function of the entire population we are interested in. To ensure the identification of the optimal dose for the population, we further assumme:

Assumption 4 *Concavity*: $\theta_T(\cdot)$ is continuous on the closed interval of $\mathcal{D}_T$ and differentiable on the open interval of $\mathcal{D}_T$, and the second derivative $\theta''_T(\cdot)$ exists throughout the domain $\mathcal{D}_T$ and $\theta''_T(d_T) < 0 \; \forall \, d_T \in \mathcal{D}_T$. This assumption guarantees the existence of the local maximum of the dose-response function within the domain $\mathcal{D}_T$.

Let $\theta_T\{g_T(\mathbf{H}_T)\}$ denote the dose-response of the outcome under regime g_T. According to Eqn (3.1), the optimal decision rule of the final stage $g_T^{opt}(\mathbf{H}_T)$ satisfies

$$E[Y^*\{g_T^{opt}(\mathbf{H}_T)\}] \geq E[Y^*\{g_T(\mathbf{H}_T)\}], \; \forall \, g_T \in \mathcal{G}_T,$$

i.e.

$$\begin{aligned} g_T^{opt}(\mathbf{H}_T) &= \arg\max_{g_T \in \mathcal{G}_T} \int_{\mathbf{H}_T} E\{Y \mid D_T = g_T(\mathbf{H}_T), \mathbf{H}_T = \mathbf{h}_T\} \mathrm{d}P(\mathbf{h}_T) \\ &= \arg\max_{g_T \in \mathcal{G}_T} \theta_T\{g_T(\mathbf{H}_T)\}. \end{aligned} \tag{3.2}$$

For the intermediate stage $t \in \{1, 2, \ldots, T-1\}$, the optimal regime g_t^{opt} is expressed in a reversed sequential order. It is not meaningful to compare the outcome of the regime of the intermediate stage when the dosage regime after the current stage are different. Therefore, we desire to compare the dosage regime of the current stage while assuming their future dosage have been optimized. We consider $\{Y^*(D_1, \ldots, D_t, g_{t+1}^{opt}, \ldots, g_T^{opt})_{D_t = d_t}\}$, for brevity $\widetilde{Y}_t^*(d_t)$, the future optimized counterfactual outcome when a patient receiving treatment d_t at stage t and all future stages dosages are under the optimal regime. Under the foregoing causal inference assumptions, the expectation of the future optimized counterfactual outcome can be identified with observed data as $E\{\widetilde{Y}_t^*(d_t)\} = E\{E(PO_t \mid D_t = d_t, \mathbf{H}_t)\}$, where the pseudo-outcome of stage t, PO_t is the calculated future optimized counterfactual outcome assuming optimal regimes are followed in all the stages after stage t. Specifically, PO_t can be defined recursively using Bellman's optimality as

$PO_t = \hat{E}\{PO_{t+1} \mid D_{t+1} = g_{t+1}^{opt}(\mathbf{H}_{t+1}), \mathbf{H}_{t+1}\}$, $t \in \{1, 2, \ldots, T-1\}$ and $PO_T = Y$ for the final stage. We can similarly denote $\theta_t(d_t) = E(\widetilde{Y}_t^*(d_t))$, and $\theta_t(\cdot)$ is the dose-response function of the entire population at stage t, and the concavity of $\theta_t(\cdot)$ should be assumed throughout the domain of $\mathcal{D}_t$. Thus, this function at stage t under the dosage regime $g_t(\mathbf{H}_t)$ is denoted as $\theta_t(g_t(\mathbf{H}_t))$. The optimal regime $g_t^{opt}(\mathbf{H}_t)$ satisfies $\theta_t(g_t^{opt}) \geq \theta_t(g_t)$, $\forall\, g_t \in \mathcal{G}_t$. In other words, the optimal regime g_t^{opt} satisfies $E(\widetilde{Y}_t^*(g_t^{opt}(\mathbf{H}_t))) \geq E(\widetilde{Y}_t^*(g_t(\mathbf{H}_t)))$, $\forall\, g_t \in \mathcal{G}_t$, i.e.,

$$\begin{aligned} g_t^{opt} &= \arg\max_{g_t \in \mathcal{G}_t} \int_{H_t} E\{PO_t \mid D_t = g_t(\mathbf{H}_t), \mathbf{H}_t = \mathbf{h}_t\} \mathrm{d}P(\mathbf{h}_t) \\ &= \arg\max_{g_t \in \mathcal{G}_t} \theta_t\{g_t(\mathbf{H}_t)\} \end{aligned} \tag{3.3}$$

3.3 KIDD-Learning with Tree-based Dose-search Algorithm

Given the observational data with continuous dosage, we propose solving Eqn (3.2) and Eqn (3.3) through KIDD-Learning to estimate the optimal individualized dynamic dose decision rule by accounting for patients' heterogeneity.

The tree-based method has been used in personalized medicine primarily (*Laber and Zhao*, 2015; *Tao et al.*, 2018; *Zhu et al.*, 2015) because it is the foundation of exploratory analysis and a classical example of a model with interpretability and predictive ability. As we also desire to obtain interpretable dosage decision rules, the tree-based method becomes an ideal candidate. At each node of a tree-based decision rule, the population is split into two subpopulations based on their medical characteristics; therefore, we can assume that patients can be classified into different sub-population based on their characteristics, and patients who fall into the same sub-population are homogeneous in terms of response to treatment doses. Thus for a given patient, the dose-response information can be inferred by borrowing information from other patients in the same group.

Therefore, the optimal tree-based dosage decision rule is estimated in the fol-

lowing steps at each treatment stage: (1) For a given split of a node, calculate the dose-response functions of the population and the two subpopulations after splitting; (2) Use a tree-based dose-search algorithm to determine whether the splitting should happen by comparing the expectation of the counterfactual dose-response outcome of the whole population. The steps are repeated iteratively until obtaining a reasonably satisfactory tree-based decision rule. In the first step, a non-parametric kernel-involved regression is applied to estimate robust dose-response functions and identify the optimal dose of the specific population. In the second step, the tree-based dose-search algorithm, which is used to evaluate the splittings and identify the tailoring variables, is able to yield an interpretable tree-based decision rule.

3.3.1 Estimating the Dose-response Function

For a given node, we denote ω as a subpopulation of the entire population Ω of the parent node. For a certain stage $t \in \{1, 2, \ldots, T\}$, we denote the causal dose-response function for a subpopulation ω at stage t as $\theta_{\omega,t}(\cdot)$, and $\theta_{\omega,t}(d_t)$, $\forall d_t \in \mathcal{D}_t$, is the dose response of the subpopulation ω when the entire subpopulation receiving dose d_t. To guarantee the optimal dose is available from $\theta_{\omega,t}(\cdot)$, we further assume that $\theta_{\omega,t}(\cdot)$ should be concave throughout the domain of $\mathcal{D}_{\omega,t}$. Therefore, the optimal dose of the subpopulation ω at stage t is denoted as $d^{opt}_{\omega,t} = \arg\max_{d_{\omega,t} \in \mathcal{D}_{\omega,t}} \theta_{\omega,t}(d_{\omega,t})$ For brevity, we suppress ω in the remaining part of this subsection when no confusion exists.

3.3.1.1 Mapping Function for the Dose-response Function

At the final stage T, let $C_T(\{D_T, \mathbf{H}_T, Y\})$ be the mapping function of the observed data such that,

$$E[C_T(\{D_T, \mathbf{H}_T, Y\}) \mid D_T = d_T] = \theta_T(d_T). \tag{3.4}$$

Given the mapping between $C_T(\{D_T, \mathbf{H}_T, Y\})$ and the dose-response function $\theta_T(d_T)$, $\theta_T(d_T)$ can be estimated by regressing $C_T(\{D_T, \mathbf{H}_T, Y\})$ on treatment d_T by using the

off-the-shelf non-parametric regression model or machine learning methods (*Kennedy et al.*, 2017).

At the intermediate stage $t \in \{1, 2, \ldots, T-1\}$, the intermediate future optimized counterfactual outcome $\tilde{Y}_t^*(d_t)$ is not directly available from the dataset. Therefore, we use the pseudo outcome as defined above to replace Y in Eqn (3.4). In particular, the mapping function before the final stage become $E[C_t(\{D_t, \mathbf{H}_t, PO_t\}) \mid D_t = d_t] = \theta_t(d_t)$.Since $PO_T = Y$, we suppress the difference between the final stage T and any intermediate stage in the following text and use $E[C_t(\{D_t, \mathbf{H}_t, PO_t\}) \mid D_t = d_t] = \theta_t(d_t)$, $t \in \{1, 2, \ldots, T\}$ for all stages. The following method works for any stage $t \in \{1, 2, \ldots, T\}$.

The mapping function $C_t(\{D_t, \mathbf{H}_t, PO_t\})$ determines the quality of the estimated dose-response function, and thus the quality of the estimated optimal decision rule. Therefore, we consider constructing a class of mapping functions $C_t(\{\mathbf{H}_t, D_t, PO_t\})$ for estimating the dose-response function $\theta_t(d_t)$. To avoid relying on the parametric assumptions or the correctly specified parametric models, we follow the semi-parametric method for continuous treatment proposed by *Kennedy et al.* (2017). We further denote that $\mathbb{P}_n$ is the empirical measure, $\pi(d_t \mid \mathbf{h}_t) = \frac{\partial}{\partial d_t} P(D_t \leq d_t \mid \mathbf{H}_t = \mathbf{h}_t)$ is the conditional treatment density given history and $\mu(\mathbf{h}_t, d_t) = E(PO_t \mid \mathbf{H}_t = \mathbf{h}_t, D_t = d_t)$ is the conditional mean pseudo outcome given covariates and treatment assignment. The mapping function can be estimated by

$$\begin{aligned} \widehat{C}_t(\{D_t, \mathbf{H}_t, PO_t\}) = \frac{PO_t - \widehat{\mu}(\mathbf{H}_t, D_t)}{\widehat{\pi}(D_T \mid \mathbf{H}_t)} \int_{\mathcal{H}_t} \widehat{\pi}(D_t \mid \mathbf{H}_t = \mathbf{h}_t) \mathrm{d}\mathbb{P}_n(\mathbf{h}_t) \\ + \int_{\mathcal{H}_t} \widehat{\mu}(\mathbf{h}_t, D_t) \mathrm{d}\mathbb{P}_n(\mathbf{h}_t). \end{aligned} \tag{3.5}$$

In particular, at the final stage T, when either of $\mu(\mathbf{H}_t, D_t)$ or $\pi(D_T \mid \mathbf{H}_T)$ is correctly specified, the estimator $C(\{D_T, \mathbf{H}_T, Y\})$ satisfies the property of double-robustness, i.e., $E[C(\{D_t, \mathbf{H}_t, Y\}) \mid D_T = d_T] = \theta_T(d_T)$. In our implementation, the conditional

treatment density model, and the conditional mean model are fitted by the method of random forest (*Breiman*, 2001).

3.3.1.2 Kernel Regression

Given such a mapping function Eqn (3.5), $\theta_t(d_t)$ can be estimated by first estimating the mapping function $\widehat{C}_t(\mathbf{H}_t, D_t, PO_t)$ and then regressing $\widehat{C}_t(\mathbf{H}_t, D_t, PO_t)$ on treatment D_t. The construction of the regression can be done by using various flexible methods. In this paper, we uses the local linear kernel regression.

The local linear kernel version of our estimator is $\widehat{\theta}_{h,t}(d_t) = \mathbf{g}_{h,d_t}(d_t)^T \widehat{\beta}_h(d_t)$, where $\mathbf{g}_{h,d_t}(z) = (1, (z - d_t)/h)^T$ and

$$\widehat{\beta}_h(d_t) = \arg\min_{\beta \in \mathbb{R}^2} \mathbb{P}_n[K_{h,d_t}(D_t)\{\widehat{C}_t(\mathbf{H}_t, D_t, PO_t) - \mathbf{g}_{h,d_t}^T(D_t)\beta\}^2],$$

where $K_{h,d_t}(z) = K\{(z - d_t)/h\}/h$ with K is a standard kernel function and h is a scalar bandwidth parameter.

The bandwidth h is essential for the kernel regression method, because too little smoothness may produce substantial variance while too much smoothness may yield large bias. In KIDD-Learning, we use the data-driven cross-validation to choose the bandwidth h for the kernel regression. Specifically, we treat the estimated $\widehat{C}_t(\{D_t, \mathbf{H}_t, PO_t\})$ as known and use the leave-one-out cross-validation to select the optimal bandwidth h_{opt}, i.e.,

$$\widehat{h}_{opt} = \arg\min_{h \in \mathcal{H}} = \sum_{i=1}^{n} \left\{ \frac{\widehat{C}_t(\{D_t, \mathbf{H}_t, PO_t\}) - \widehat{\theta}_{t,h}(D_{t,i})}{1 - \widehat{W}_h(D_{t,i})} \right\}^2,$$

where $\widehat{\theta}_{t,h}(D_{t,i})$ is the estimated dose response for patient i when the bandwidth equals to h, and $\widehat{W}_h(D_{t,i}) = (1,0)\mathbb{P}_n\{\mathbf{g}_{hd_t}(d_t)K_{h,d_{t,i}(D_t)}\mathbf{g}_{h,d_t}(d_t)^T\}^{-1}(1,0)^T h^{-1}K(0)$ is the ith diagonal of the hat matrix. We expect that this approach of selecting optimal bandwidth h_{opt} is asymptotically equivalent to using the oracle bandwidth selector

(*Kennedy et al.*, 2017).

3.3.2 Tree-based Dose-search Algorithm for KIDD-Learning

The algorithm of dose search is a modified method of the tree-based reinforcement learning method (*Laber and Zhao*, 2015), which uses the classification and regression tree (CART) proposed by *Breiman et al.* (1984). CART explores the relationship between the directly observed or given classification label and the covariates, and then builds a decision tree with the purest child nodes, which means having the lowest mis-classification rate among all possible splits. In this dose-search problem, the optimal dose, i.e., the label, is unknown and only available indirectly through the counterfactual outcome. Therefore, to make use of this indirect information, we use the expected counterfactual outcome as the purity measure, and build the tree-based decision rules by maximizing the expected counterfactual outcome. Since the tree-based dose-search algorithm of KIDD-learning involves infinite dose options within the given dose range, the purity measure should be modified accordingly using the estimation the dose-response function.

3.3.3 Implementation of KIDD-Learning

3.3.3.1 Purity Measure

A measure of node purity is used to facilitate the recursive splitting procedures when growing a decision tree. For a given partition ω and ω^c of a node Ω, we evaluate the dose-response function separately for $\mathbf{H}_t \in \omega$ as $\theta_{\omega,t}(d_{\omega,t})$ and for $\mathbf{H}_t \in \omega^c$ as $\theta_{\omega^c,t}(d_{\omega^c,t})$. We first define the purity measure of stage t as

$$\mathcal{P}_t(\Omega, \omega) = \widehat{\theta}_{\omega,t}(d_{\omega,t})\mathbb{P}_n[I(\mathbf{H}_t \in \omega)] + \widehat{\theta}_{\omega^c,t}(d_{\omega^c,t})\mathbb{P}_n[I(\mathbf{H}_t \in \omega^c)],$$

where $\mathbb{P}_n$ is the empirical expectation operator. Under the causal assumptions, $\widehat{\theta}_{\omega,t}(d_{\omega,t})$ is the estimator of the expected dose response of subpopulation ω if all the subjects in ω takes dose $d_{\omega,t}$. However, comparing the purity by enumerating all possible $d_{\omega,t}$ and $d_{\omega^c,t}$ is not efficient to find the best partition. Instead, we first maximize each partition's purity by plugging in the optimal dose of each partition ω and ω^c, and then accept the partition with the highest purity as ω^{opt}. Based on the assumption of the concavity of the dose-response function, for a given partition ω and ω^c, the purity measure can be re-written as

$$\mathcal{P}_t(\Omega, \omega) = \widehat{\theta}_{\omega,t}(d^{opt}_{\omega,t})\mathbb{P}_n[I(\mathbf{H}_t \in \omega)] + \widehat{\theta}_{\omega^c,t}(d^{opt}_{\omega^c,t})\mathbb{P}_n[I(\mathbf{H}_t \in \omega^c)], \tag{3.6}$$

where $d^{opt}_{\omega,t} = \arg\max_{d_{\omega,t}\in\mathcal{D}_{\omega,t}} \widehat{\theta}_{\omega,t}(d_{\omega,t})$ and $d^{opt}_{\omega^c,t} = \arg\max_{d_{\omega^c,t}\in\mathcal{D}_{\omega^c,t}} \widehat{\theta}_{\omega^c,t}(d_{\omega^c,t})$.

3.3.3.2 Recursive Partitioning

When maximizing the purity measure $\mathcal{P}_t(\Omega, \omega)$ as Eqn (3.6), a split ω can be either a category level of a categorical covariate or values of continuous covariates that are not larger than the threshold. The best partition ω^{opt}, which leads to the dose assignments $d^{opt}_{\omega,t}$ and $d^{opt}_{\omega^c,t}$, should maximize the improvement of $\mathcal{P}_t(\Omega - \omega) - \mathcal{P}_t(\Omega)$, where $\mathcal{P}_t(\Omega)$ means no splitting; i.e., all the subjects in Ω are assigned the same dose. It is obvious that $\mathcal{P}_t(\Omega - \omega) - \mathcal{P}_t(\Omega) \geq 0$. To provide a practically meaningful split, the improvement of purity should exceed a positive threshold λ . In addition, to avoid overfitting and prune a feasible decision tree, we define the minimal node size n_0, the minimum number of subjects of a terminal node, and $depth_{\max}$, the maximal number of edges from the node to a tree's top root node. If n_0 is too small or $depth_{\max}$ is too large, we will end up with an overfitted, overly precise dosage decision rule. Given the hyper-parameters, λ, $depth_{\max}$ and n_0, the following Algorithm 2 is implemented as the stopping rules to split each node and stop while the rules are violated. Note

here, n_Ω is the total number of subjects in the parent node Ω.

Algorithm 2 Stopping Rules

if the current tree depth reaches the pre-specified $depth_{\max}$ **then**

Do not split the node

else

Calculate the best split by

$$\widehat{\omega}^{opt} = arg\max_{\omega} \left[\mathcal{P}_t(\Omega, \omega) : \min\{n_\Omega \mathbb{P}_n I(\mathbf{H}_t \in \omega), n_\Omega \mathbb{P}_n I(\mathbf{H}_t \in \omega^c)\} \geq n_0\right].$$

if the maximum purity improvement $\mathcal{P}_t(\Omega, \widehat{\omega}^{opt}) - \mathcal{P}_t(\Omega) < \lambda$ **then**

Do not split the node

else

Split Ω into $\widehat{\omega}^{opt}$ and $\widehat{\omega}^{opt^c}$

end if

end if

3.3.3.3 Implementation

KIDD-Learning is implemented by a backward induction in a reversed sequential order. At each stage, the optimal dosage decision is made by recursively evaluating the dose-response function $\theta_t(d_t)$ of each stage. Before determining the dose-response function of each stage t, we have to calculate the pseudo outcome PO_t, which is the replacement of the outcome Y in the intermediate stages. At the final stage $t = T$, $PO_T = Y$, and thus PO_T can be used directly from the observed outcome. In the stage $t \in \{1, \ldots, T-1\}$, the pseudo-outcome PO_t requires that the estimated dosage decision rules of the future stages have been already optimized, as $PO_t = \hat{E}\{PO_{t+1} \mid D_{t+1} = g_{t+1}^{opt}(\mathbf{H}_{t+1}), \mathbf{H}_{t+1}\}$. We further assume the dosage effect of the intermediate outcome Y_t is cumulatively carried forward to the final outcome Y. To reduce the accumulated bias of each stage, we calculate the pseudo outcome PO_t by

using actual observed outcomes at stage t, plus the predicted pseudo outcome gain due to the optimized future dose assignment of the stages after t (*Huang et al.*, 2015); that is

$$PO_t = Y + \sum_{j=t+1}^{T} \{E(PO_j \mid \mathbf{H}_j, D_j = g_j^{opt}(\mathbf{H}_j)) - E(PO_j \mid \mathbf{H}_j, D_j = d_j)\},$$

where $PO_T = Y$ and the conditional outcome of PO_t is calculated by using the off-shelf machine-learning method "*eXtreme Gradient Boosting*" (*Chen and Guestrin*, 2016). In practice, the intermediate outcomes Y_t can be different measures of a patient's disease status, such as toxicity or efficacy, due to different priorities of treatment during the progression of the disease. Therefore, we should standardize the scale of different intermediate measurements before aggregating them from different stages into the final outcome. In some particular stage where the intermediate outcome cannot be directly observed but the effect of the outcome is carried forward, we can specify these outcomes as zero. The implementation of KIDD-Learning is outlined in Algorithm 3, where $m \in \{1, 2, \ldots, 2^{depth_{\max}} - 1\}$ is the index of the node.

3.4 Extension to Survival Outcome

A time-to-event outcome, such as overall survival or local progression, is commonly used in oncology studies to evaluate the treatment performance. However, a patient may have been censored from the study before the event happens, and the observed censored time for the patient cannot be directly used as a continuous outcome. Therefore, we extend the proposed method from using continuous outcomes to using survival outcomes, following the recursive imputed survival time (RIST) method proposed by *Zhu and Kosorok* (2012). We modify RIST to transform the actual censored time into the expected conditional survival time, and then use it as a continuous outcome for KIDD-Learning.

Algorithm 3 Implementation of KIDD-Learning

Result: $\mathbf{g}^{opt} = g_1^{opt}, \ldots, g_T^{opt}$

Set $t = T$ and $PO_T = Y$

while $t \geq 1$ **do**

Set $m = 1$ at root node $\Omega_{t,m}$

while $m \leq 2^{depth_{\max}} - 1$ **do**

Use subjects with history $\mathbf{H}_t \in \Omega_{t,m}$

if Stop at the *Stopping rules* at node $\Omega_{t,m}$ **then**

Assign a single best treatment $d_t^{opt} = \arg\max_{d_t \in \mathcal{D}_t} \theta_{t,\Omega_{t,m}}(d_t)$ to all subject in $\Omega_{t,m}$

else

Split $\Omega_{t,m}$ into child nodes $\Omega_{t,2m}$ and $\Omega_{t,2m+1}$ by $\widehat{\omega}^{opt}$

end if

$m = m + 1$

end while

$t = t - 1$

Set $PO_t = Y + \sum_{j=t+1}^{T} \{E(PO_j \mid \mathbf{H}_j, D_j = g_j^{opt}(\mathbf{H}_j)) - E(PO_j \mid \mathbf{H}_j, D_j = d_j)\}$,

end while

For brevity, we suppress the stage subscript t in the following derivation when no confusion exists. Let T_i be the event (death) time for of the i^{th} patient, let C_i be the censoring time for the i^{th} patient, and let $\delta_i = I(T_i \leq C_i)$ be the non-censoring indicator, 1 as event (death) and 0 as censored. We further define the observed survival time $R_i = \min(T_i, C_i)$, and let X_i be the patient's history and D_i be the dose assignment of the patient i. The observed data consist n i.i.d. $\{R_i, \delta_i, X_i, D_i\}_{i=1}^n$. Since the event time is not always observed, we use a replacement of the observed event time while maintaining a similar outcome. Therefore, we denote Y_i as the outcome, which is defined as the expectation of survival time T_i conditional on the covariates X_i, and D_i, censoring status δ_i and the observed survival time R_i. We further assume that the censoring is independent of the survival time conditional on the covariates (*Cui et al.*, 2017). We also assume that there is a maximum length of follow-up time τ. Since the survival time of event patients, i.e., $\delta_i = 1$, is known, we focus on estimating the conditional expectation of T which is truncated at τ as,

$$\begin{aligned} Y_i &= E(T_i \mid X_i, D_i, R_i, \delta) \\ &= I(\delta_i = 1)R_i + I(\delta_i = 0)E(T_i \mid X_i, D_i, T_i > R_i) \end{aligned}$$

Next, we can calculate the conditional survival time for the censored patients by taking the integral of the conditional survival function, and the conditional survival function can be estimated by fitting survival decision trees (*Ishwaran and Lu*, 2019).

To have a more stable conditional survival function, we repeat the survival tree splitting for K times and then average over all K trees. For each terminal node in k^{th} tree, we calculate a Kaplan-Meier estimate of the survival function within each node, which is denoted by $\widehat{S}_k(t)$. Every patient will only fall into one terminal node for each fitted decision tree. Therefore, we denote the single-tree survival function by $\widehat{S}_k^i$ for the i^{th} subject. By averaging K trees, we have the "forest-level" survival function $\widehat{S}_i(t) = \frac{1}{K}\sum_{k=1}^K \widehat{S}_k^i(t)$. Given a subject i who is censored at time R_i, we can

approximate the conditional probability of survival, $Pr(T_i > t \mid X_i, D_i, T_i > R_i,)$, by

$$Pr(T_i > t \mid X_i, D_i, T_i > R_i,) = \begin{cases} 1 & \text{if } t \in [0, R_i] \\ \widehat{S}_i(t)/\widehat{S}_i(R_i) & \text{if } t \in (R_i, \tau], \end{cases}$$

where τ is the pre-specified maximal length of follow-up time of the study.

3.5 Simulation Studies

We conduct simulation studies to investigate the performance of KIDD-Learning under different scenarios. At first, to facilitate the comparison with existing methods, in particular CZK (*Chen et al.*, 2016), Q-learning regression-based SVR, and LZ (*Laber and Zhao*, 2015), we consider single-stage scenarios and two-stage scenarios. For each scenario, we consider sample sizes of 500, 800 or 1000 for the training dataset and evaluate the performance on a test dataset with a sample size N=1000, and replicate them 500 times.

The training datasets are used to estimate the optimal dosage regime and then predict the optimal dosage decision in a separate test dataset, where the underlying true optimal dosage regime is already known. The *opt*% shows the average proportion of subjects correctly classified to their true optimal dose according to the estimated optimal dosage regime. The expectation of the estimated counterfactual dose-response outcome $\widehat{E}(Y^*(\widehat{g}^{opt}))$, i.e., the expected dose response under the estimated dosage regime $\widehat{g}^{opt}$, is also used to evaluate the performance. This expected value is calculated by plugging the estimated optimal dose into the true pre-specified underlying value function. Since the results of *opt*% and $\widehat{E}(Y^*(\widehat{g}^{opt}))$ are not normally distributed (*Tao et al.*, 2018), the results shown in Table 3.1 and 3.2 are medians and the interquartile range from 25^{th} quartile to 75^{th} quartile.

3.5.1 Single-Stage Scenarios

We consider a single-stage case with a continuous dose and implement KIDD-Learning in the training dataset, of which the sample size is 500, 800 or 1000. We first generate three baseline covariates X_1, X_2, X_3 according to $N(0,1)$. The treatment D_1 is set within the range of $\mathcal{D}_1 = [0,1]$, and we generate it from $Beta(\alpha, 1-\alpha)$ with $\alpha = 1/\{1+\exp(0.3X_1 + 0.2X_2 + 0.1X_3)\}$, which makes D_1 depend on the observed covariates.

The continuous outcome Y_1 has the form of $Y_1 = m(D_1) + \beta\mathbf{H_1} + \epsilon_1$, where $\beta\mathbf{H_1} = 0.5X_1 + 0.3X_2 + 0.7X_3$ and ϵ_1 is an independent standard normal variate that follows $N(0,1)$. In addition, $m(d_1) \propto I(g_1^{opt} = 0.5)F_{24,24}(d_1) + I(g_1^{opt} = 0.2)F_{6.75,24}(d_1) + I(g_1^{opt} = 0.8)F_{24,6.75}(d_1)$, where $F_{p,q}(d_1)$ is a unimodal function within the domain to ensure the existence of the maximal dose response. Here, $F_{p,q}(d_1) = (d_1)_1^{p-1}(1-d_1)^{q-1}\Gamma(p+q)/\{\Gamma(p)\Gamma(q)\}$ is the probability density function of $Beta(p,q)$.

We consider two forms of the underlying optimal dosage assignment, one is a tree-type dosage regime and the other one is a non-tree-type regime, to study the impact of underlying model specification has on the performance of KIDD-Learning. The underlying tree-type is consistent with the dosage assignment model when using the proposed tree-based dose-search algorithm, while the underlying model is "mis-specified" for the non-tree-type regime. In particular, the underlying optimal dose for the tree-type dosage regime is specified as $g_1^{opt}(\mathbf{H}_1) = 0.8I(X_1 > 0.4) + 0.2I(X_1 \leq 0.4)I(X_2 \leq 0) + 0.5I(X_1 \leq 0.4)I(X_2 > 0)$. The optimal dosage regime for the non tree-type is specified as $g_1^{opt}(\mathbf{H}_1) = 0.8I(X_1 + X_2 > 0.6) + 0.2I(X_1 + X_2 \leq 0.6)I(X_2 \leq -0.3) + 0.5I(X_1 + X_2 \leq 0.6)I(X_2 > -0.3)$.

For comparison, three methods, CZK, SVR, and LZ, are considered. CZK, proposed by *Chen et al.* (2016), is a non-linear version of the modified outcome weighted learning method to estimate the individualized dosage regime, a special case of DTR. It uses a non-convex loss function for optimization and provides a difference of convex

functions to solve the optimization problem. SVR includes a two-step procedure similar to Q-learning: in the first step, the non-linear relationship between the outcomes and the dosage decision, and relationship between covariates and dose are evaluated by a Gaussian kernel. The parameters of SVR are tuned by five-fold cross-validation. In the second step, since the closed-form expression of the optimal dosage regime is not available, we choose 200 equally distributed grids within the dose interval $(0, 1)$ and find the optimal dose, and thus the optimal dosage regime. In addition, we use the a non-parametric method for CZK and SVR, as presented in *Chen et al.* (2016), to estimate the density of having a certain dose conditional on the covariates. LZ is another tree-based reinforcement learning method that uses CART to estimate the optimal DTR for continuous dose. Comparing with KIDD-Learning, LZ calculates the purity measure by using the kernel density and evaluates the purities of given doses to yield the optimal dosage regime. These comparison are implemented using R packages *"kernlab"* (*Karatzoglou et al.*, 2018), *"SVMW"* (*Chen et al.*, 2016) and *"np"* (*Hayfield and Racine*, 2008).

Table 3.1 summarizes the result for the single stage scenario. It shows that KIDD-Learning outperforms CZK, SVR and LZ in all settings. Specifically, when sample size N=1000, KIDD-Learning correctly estimates 97.3% subjects into their optimal dosage regime, and the median of the estimated estimated mean counterfactual outcome is 9.72, which is very close to the pre-specified true value 10. When the sample size decreases, the performance of KIDD-Learning decreases as expected. When the sample size N=500, the performance decreases but is still superior to CZK, SVR and LZ with the median of the estimated counterfactual outcome equals 9.06 with 90.65% of the subjects correctly classified. CZK uses more information from the observations who have higher observed outcome; if the received doses of a large part of the population are close to their optimal doses, the performance of CZK should be satisfactory. However, it is not the case in our simulation study and not common in

the observational study either. Therefore, the compromised performance of CZK is to be expected (*Chen et al.*, 2016).

We further consider a setting that mimics a situation when the tailoring variables that interact with the dose assignment are not clear to the researchers. Specifically, it is a setting with 7 additional baseline variables,$X_4, \ldots X_{10}$, simulated independently from $N(0,1)$, and the outcome $Y_1 = m(D_1) + \beta\mathbf{H_1} + \epsilon_1$, follows the previous setting, where $\beta\mathbf{H_1} = 0.5X_1 + 0.3X_2 + 0.7X_3 + \sum_{i=4}^{10} 0.2X_i$. In this setting, we intend to see how the noise interference may affect the performance of the four methods under different sample sizes and different underlying optimal dosage regimes.

The results of simulation study including more noise baseline covariates are shown in Table 3.2. In general, KIDD-Learning has the highest performance of the methods compared. Although all methods in all cases have worse performance than the results shown in Table 3.1, the regression-based SVR is the most sensitive to the increase of the dimension of noise covariates. When the sample size is 1000 and under the pre-specified tree-type dynamic dosage regime, the median of estimated expectation of counterfactual outcome $E\{Y^*(\widehat{g}^{opt})\}$ of SVR is 3.10, which is almost half of that in the case with fewer noise variables. Compared with Table 3.1, KIDD-Learning has a decreased performance in *opt*% and $\widehat{E}(Y^*(\widehat{g}^{opt}))$ in general, but the performance is still reasonably satisfactory when we have a sufficient sample size, i.e., N=1000 or N=800. Specifically, the median of expectation of counterfactual outcome when N=1000 is 9.60, and *opt*% is 96.2%, which is only decreased by 1% from its counterpart in Table 3.1. In general, the performance gets worse when the sample size decreases, but the performance of KIDD-Learning is still much better than that of CZK,SVR and LZ.

Table 3.1: Simulation results for a single-stage scenario with three baseline covariates (500 replications, N=500, 800 or 1000). The tree-type and the non tree-type are two different pre-specified optimal dosage regimes. $E\{Y^*(\widehat{g}^{opt})\} = 10$

		Tree-type		Non Tree-type	
Sample Size	Method	$E\{Y^*(\widehat{g}^{opt})\}$	$opt\%$	$E\{Y^*(\widehat{g}^{opt})\}$	$opt\%$
500	KIDD-L	9.06(7.65, 9.60)	90.65(75.77, 96.12)	7.32(6.41, 8.08)	72.70(63.90, 80.80)
	CZK	4.76(4.55, 4.98)	28.40(26.50, 30.63)	4.93(4.68, 5.15)	28.80(26.20, 31.30)
	SVR	6.72(6.48, 6.94)	58.40(55.80, 60.90)	6.19(5.97, 6.43)	55.40(52.00, 58.20)
	LZ	7.33(6.40, 8.09)	54.55(38.00, 67.80)	6.26(5.42, 6.96)	45.00(30.28, 56.20)
800	KIDD-L	9.63(9.37, 9.78)	96.30(93.68, 98.10)	8.14(7.48, 8.41)	81.40(74.47, 84.20)
	CZK	5.12(4.93, 5.28)	31.40(29.30, 33.23)	5.29(5.08, 5.47)	31.40(29.20, 33.60)
	SVR	7.30(7.13, 7.46)	64.80(62.30, 66.80)	6.79(6.59, 6.97)	60.90(58.60, 63.30)
	LZ	8.67(8.19, 9.02)	77.10(68.38, 84.10)	7.53(6.95, 7.95)	66.05(55.48, 73.50)
1000	KIDD-L	9.72(9.51, 9.84)	97.30(95.20, 98.53)	8.27(7.87, 8.49)	82.55(78.58, 84.92)
	CZK	5.29(5.12, 5.48)	33.20(30.90, 35.00)	5.43(5.27, 5.60)	32.55(30.48, 34.62)
	SVR	7.55(7.37, 7.70)	67.20(65.10, 69.40)	7.04(6.84, 7.22)	63.70(61.70, 65.62)
	LZ	9.10(8.72, 9.33)	84.70(77.90, 88.95)	7.95(7.55, 8.28)	73.50(65.58, 79.00)

a. $opt\%$ shows the median and the interquartile range of the percentage of patient exactly correctly classified to the optimal true dose.
b. $\widehat{E}\{Y^*(\widehat{g}^{opt})\}$ shows the median and its interquartile range with the estimated mean counterfactual outcome obtained using the true outcome model and the estimated optimal dynamic dosage regime.

Table 3.2: Simulation results for single-stage scenarios with ten baseline covariates (500 replications, N=500, 800 or 1000). The tree-type and the non tree-type are two different pre-specified optimal dosage regimes. $E\{Y^*(\widehat{g}^{opt})\} = 10$

		Tree-type		Non Tree-type	
Sample Size	Method	$E\{Y^*(\widehat{g}^{opt})\}$	$opt\%$	$E\{Y^*(\widehat{g}^{opt})\}$	$opt\%$
500	KIDD-L	6.35(3.75, 8.77)	62.65(37.35, 87.10)	6.14(3.94, 6.64)	61.10(39.05, 66.32)
	CZK	3.21(3.07, 3.36)	17.20(16.10, 18.30)	3.37(3.21, 3.57)	18.30(16.90, 19.42)
	SVR	3.10(2.92, 3.27)	16.40(15.20, 17.70)	3.19(2.99, 3.36)	17.80(16.40, 19.20)
	LZ	6.87(6.19, 7.53)	45.25(33.85, 57.00)	5.84(5.12, 6.52)	38.20(26.60, 49.50)
800	KIDD-L	9.33(7.76, 9.75)	93.20(76.35, 97.70)	6.79(6.29, 7.84)	67.45(62.60, 78.45)
	CZK	3.37(3.25, 3.49)	18.25(17.10, 19.30)	3.56(3.41, 3.70)	19.50(18.40, 20.60)
	SVR	3.33(3.17, 3.48)	17.90(16.67, 19.20)	3.40(3.24, 3.54)	18.85(17.67, 20.20)
	LZ	8.35(7.72, 8.76)	70.95(59.20, 78.90)	7.21(6.68, 7.64)	59.50(49.82, 67.33)
1000	KIDD-L	9.60(9.22, 9.80)	96.20(92.18, 98.20)	7.58(6.57, 8.21)	75.70(65.60, 82.12)
	CZK	3.43(3.32, 3.54)	18.80(17.80, 19.70)	3.60(3.48, 3.73)	19.80(18.70, 21.00)
	SVR	3.47(3.34, 3.65)	18.90(17.70, 20.40)	3.55(3.41, 3.70)	20.05(18.67, 21.42)
	LZ	8.80(8.38, 9.05)	79.35(72.10, 84.20)	7.61(7.22, 7.96)	66.90(59.80, 72.85)

a. $opt\%$ shows the median and the interquartile range of the percentage of patient exactly correctly classified to the optimal true dose.
b. $\widehat{E}\{Y^*(\widehat{g}^{opt})\}$ shows the median and its interquartile range with the estimated mean counterfactual outcome obtained using the true outcome model and the estimated optimal dynamic dosage regime.

3.5.2 Two-Stage Scenarios

To mimic the multiple-stage dynamic dosage regime, data under a two-stage continuous dynamic dosage regime are generated and evaluated at sample sizes of 500, 800 or 1000. The outcome we are interested in is the sum of the intermediate outcomes of each stage, i.e., $Y = Y_1 + Y_2$. Similar as the single stage scenarios, we also consider the underlying true optimal dosage regime as tree-type or non tree-type.

To help make decisions in the second stage, we generate two time-varying biomarkers X_4 and X_5, in addition to three baseline covariates X_1, X_2, X_3 as the setting of the single-stage scenario; $X_1, X_2, \ldots, X_5$ are simulated independently from $N(0,1)$. For the first stage, we generate D_1 from $Beta(\alpha, 1-\alpha)$ with $\alpha = 1/\{1+\exp(0.3X_1 + 0.2X_2 + 0.1X_3)\}$. The continuous outcome of the first stage Y_1 has the form as $Y_1 = m_1(D_1) + \beta\mathbf{H}_1 + \epsilon_1$, where $\beta\mathbf{H}_1 = 0.5X_1 + 0.3X_2 + 0.7X_3$ and $m_1(d_1) \propto I(g_1^{opt} = 0.5)F_{24,24}(d_1) + I(g_1^{opt} = 0.2)F_{6.75,24}(d_1) + I(g_1^{opt} = 0.8)F_{24,6.75}(d_1)$. We also assume the independent standard normal variate $\epsilon_1 \sim N(0,1)$. Specifically, the underlying optimal dosage regime of stage 1 for the tree-type dosage regime is specified as $g_1^{opt}(\mathbf{H}_1) = 0.8I(X_1 > 0.4) + 0.2I(X_1 \leq 0.4)I(X_2 \leq 0) + 0.5I(X_1 \leq 0.4)I(X_2 > 0)$. The optimal dosage regime for the non tree-type is specified as $g_1^{opt}(\mathbf{H}_1) = 0.8I(X_1+X_2 > 0.6)+0.2I(X_1+X_2 \leq 0.6)I(X_2 \leq -0.3)+0.5I(X_1+X_2 \leq 0.6)I(X_2 > -0.3)$. At the second stage, the treatment assignment D_2 is also generated from $Beta(\alpha, 1-\alpha)$ with $\alpha = 1/\{1+\exp(0.3X_5+0.2X_5+0.1X_3)\}$. The continuous outcome $Y_2 = m_2(D_2)+\beta\mathbf{H}_2+\epsilon_2$, where $\beta\mathbf{H}_2 = 0.3X_1+0.2X_2+0.5X_3+0.6X_4+0.2X_5$ and $m_2(d_2) \propto I(g_2^{opt} = 0.5)F_{24,24}(d_2) + I(g_2^{opt} = 0.2)F_{6.75,24}(d_2) + I(g_2^{opt} = 0.8)F_{24,6.75}(d_1)$. Again, we assume $\epsilon_2 \sim N(0,1)$. The underlying optimal dosage regime of stage 2 for the tree-type dosage regime is specified as $g_2^{opt}(\mathbf{H}_2) = 0.8I(Y_1 > 1) + 0.2I(Y_1 \leq 1)I(X_4 \leq -0.1) + 0.5I(Y_1 \leq 1)I(X_4 > -0.1)$. The optimal dosage regime for the non tree-type is specified as $g_1^{opt}(\mathbf{H}_1) = 0.8I(Y_1 + X_5 > 1.3) + 0.2I(Y_1 + X_5 \leq 1.3)I(X_4 \leq 0) + 0.5I(Y_1 + X_5 \leq 1.3)I(X_4 > 0)$

Since the performances of CZK from the single-stage scenarios is consistently inferior to other methods, we only modify the regression-based SVR and LZ for comparison in the two-stage scenario. Specifically, to use SVR in this two-stage scenario, we apply the multiple-stage Q-learning and use support vector regression as the regression model. The search for optimal dosage regime is similar to that of the single-stage scenarios.

Results for different sample sizes and underlying optimal dosage regimes are shown in Table 3.3. Given a tree-type underlying dynamic dosage regime with a sample size N=1000, the median of the estimated mean counterfactual outcome of KIDD-Learning is 19.56, which is fairly close to the optimal value of 20. For both stages, 96.0% of the patients are correctly assigned to their optimal dosage regime, with 96.4% for stage 1 and 99.9% for stage 2. In contrast, SVR only has 7.81 in the median of the estimated mean counterfactual outcome, and the *opt*% is only 9.6% for both stages, and LZ is better than SVR with $\widehat{E}(Y^*(\widehat{g}^{opt})) = 14.77$ and *opt*%=22.25%. A similar superiority of KIDD-Learning is also found under different types of underlying dynamic dosage regimes and different sample sizes. Moreover, when the sample size decreases, the performance of all methods is compromised. Specifically, the convergence rate of the dose-response function using kernel regression is much slower than that of the parametric case; therefore, the relatively worse performance with a smaller sample size is to be expected. In particular, when the underlying true dynamic dosage regime is correctly specified, i.e., under the tree-type setting, the performance of KIDD-Learning with sample size N=500 is still reasonably satisfactory, with $\widehat{E}(Y^*(\widehat{g}^{opt})) =$ 17.51 and 77.2% of patients correctly assigned to their optimal dosage regimes.

Table 3.3: Simulation results for two-stage scenarios with three baseline covariates and two time-varying covariates (500 replications, N=500, 800 or 1000). Tree-type and non tree-type are two different pre-specified underlying optimal dosage regimes. $E\{Y^*(\widehat{g}^{opt})\} = 20$.

	Sample Size	Method	$\widehat{E}\{Y^*(\widehat{g}^{opt})\}$	Stage 1 *opt*%	Stage 2 *opt*%	Overall *opt*%
Tree-type	500	KIDD-L	17.51(14.69, 19.21)	81.50(65.00, 94.32)	97.70(83.57, 99.60)	77.25(53.63, 93.33)
		SVR	5.87(5.31, 6.46)	23.60(19.60, 27.60)	20.75(17.10, 24.22)	4.50(2.70, 7.00)
		LZ	12.33(11.08, 13.66)	30.85(18.48, 45.70)	46.85(32.10, 62.80)	11.15(2.15, 22.35)
	800	KIDD-L	19.41(18.67, 19.72)	95.10(89.07, 97.90)	99.80(99.30, 100.00)	94.75(86.15, 97.80)
		SVR	6.96(6.38, 7.54)	29.50(24.90, 33.40)	24.00(20.60, 27.20)	7.80(5.40, 10.43)
		LZ	14.07(13.00, 15.16)	32.40(18.15, 45.73)	68.10(53.92, 80.23)	17.45(7.38, 30.22)
	1000	KIDD-L	19.56(19.12, 19.77)	96.40(92.60, 98.32)	99.90(99.70, 100.00)	96.05(91.40, 98.20)
		SVR	7.81(7.30, 8.35)	33.20(28.25, 37.60)	29.80(26.60, 32.90)	9.65(7.20, 12.72)
		LZ	14.77(14.01, 15.74)	33.45(21.45, 47.65)	76.65(61.88, 85.70)	22.25(10.60, 33.75)
Non Tree-Type	500	KIDD-L	13.88(11.17, 15.91)	66.10(56.27, 76.82)	73.35(52.05, 88.73)	52.15(27.05, 68.80)
		SVR	5.71(5.21, 6.28)	24.70(21.28, 28.60)	17.50(14.10, 21.42)	4.40(2.60, 6.80)
		LZ	11.48(9.96, 12.85)	29.40(14.75, 41.35)	48.60(33.85, 63.02)	12.80(2.12, 24.65)
	800	KIDD-L	16.51(14.47, 17.52)	77.10(67.20, 82.22)	91.90(69.62, 95.40)	69.40(50.20, 80.18)
		SVR	6.67(6.17, 7.29)	29.55(25.50, 33.02)	20.75(17.30, 24.30)	7.00(4.80, 9.50)
		LZ	12.62(11.69, 13.56)	21.75(8.47, 37.78)	66.75(54.40, 76.20)	12.25(3.70, 24.30)
	1000	KIDD-L	17.16(15.34, 17.85)	80.65(73.77, 83.70)	94.40(78.18, 96.10)	75.40(58.85, 82.80)
		SVR	7.58(7.06, 8.15)	34.60(29.67, 38.40)	26.10(23.20, 29.30)	8.90(6.60, 11.23)
		LZ	12.90(12.00, 13.86)	17.20(5.78, 34.70)	71.55(62.08, 80.90)	10.75(3.70, 23.52)

a. *opt*% shows the empirical mean of the percentage of subjects correctly predicted their optimal dosage regimes.
b. $\widehat{E}(Y^*\{\widehat{g}^{opt}\})$ shows the median and the interquartile range of the estimated mean counterfactual outcome obtained using the true outcome model and the estimated optimal dynamic dosage regime.

3.6 Application to Liver Cancer Adaptive Stereotactic Body Radiation Therapy Data from Michigan Medicine

In radiation therapy, higher doses of radiation therapy (RT) are often associated with improved tumor control but also increased risk of toxicity. At the same time, by splitting the treatment into multiple stages, with a break in between, the radiated normal cells may recover better from the treatment and reduce the risk of toxicity (*Jackson et al.*, 2019). Thus, selection of the optimal dose requires a tradeoff between efficacy and toxicity when planning the optimal treatment dosage, accounting for individual patient characteristics. Patients vary in terms of their sensitivity to radiation therapy (*Lawrence et al.*, 1995), so splitting an RT plan into two stages provides the opportunity to identify the more sensitive patients based on early measures following the first stage of treatment. However, to our knowledge, there are no existing applications of a dynamic treatment regime framework in the setting of radiation oncology.

We apply the proposed KIDD-Learning method to a Michigan Medicine liver cancer dataset comprised of several clinical trials using adaptive SBRT (*Feng et al.*, 2018, 2013). In these studies, patients were treated with an initial RT plan followed by an assessment of liver function after a one-month break. After the treatment break, additional doses were given depending upon the change in measures of liver function. The dataset we used consisted of 202 patients treated with a range of doses during stage 1 due to physician preference, tumor size and tumor location. 84 patients stopped RT after stage 1 with no further dose, 21 patients received a lower than planned dose for the second stage, and 97 continued with the planned dose for the second stage. Such heterogeneity of treatment allowed us to apply KIDD-Learning to estimate a optimal personalized dosage regime; in other words, we consider what dose should be given for patients for the first stage using baseline covariates and what

is the optimal radiation dose for the second stage, using both baseline covariates and updated covariates observed after stage 1.

Patients' RT tumor dose ranged from 29 Gy to 119 Gy for the first stage, and from 0 Gy (no further RT) to 71 Gy for the second stage. Patients' baseline characteristics, including age, gender, cancer history, tumor size, gross tumor volume, mean liver dose, and other biomarkers (ALBI raw score, Child-Pugh score, ECOG.PS score, AST, ALT and Alkphos), were collected. The key concept underlying the two-stage dosage regime is that the full planned course of treatment may be too toxic for some patients. Thus the plan was to give 60% of the total planned dose during stage 1 and then, based on the resulting change in liver function, possibly give an additional dose, but no greater than initially planned. Exactly how the stage 2 dose was selected varied across patients.

To consider the trade-off between efficacy and toxicity, overall survival becomes an appropriate outcome to be maximized when estimating the optimal dosage regime, because severe toxicity damages the normal liver function, which then impacts the long-term life expectancy; on the other hand, better efficacy on the cancer cells avoids patients suffering from cancer progression, leading to a longer and better life. Overall survival also matches our method as we assume higher outcome values are preferable. In particular, when using the overall survival as the final outcome, we are assuming all intermediate outcomes as zero and only the overall survival after the final stage can be used as the final outcome. We use the RIST (*Zhu and Kosorok*, 2012) as described in section 3.4 to impute the conditional survival time for the patients who were censored from the study before their death. This conditional time is truncated at 3821 days, which is the longest observing time of this adaptive SBRT study.

The estimated optimal dosage regime of this two-stage adaptive SBRT is shown in Figure 3.1. In particular, after consulting with radiologists, the dose recommendation of the estimated optimal regime is one of the pre-specified candidate dose levels that

uses 10-quantiles from the observed dose, instead of an overly precise dose on a continuum. According to the estimated optimal dosage regime, at the first treatment stage, patients whose ALT baseline was larger than 58 should be given 29 Gy. Patients whose ALT baseline was larger than 50 but less than 58 were recommended for 55 Gy at the first stage. This splitting rule that gives patients with a worse baseline liver condition a lower dose has been suggested in the literature (*Lawrence et al.*, 1995). Moreover, for patients with ALT baseline less than 50, but AST baseline larger than 39, 55 Gy should also be given. For the remaining patients with lower ALT (<50) and lower AST (<39), their dose recommendations were based on their tumor size; i.e., a larger tumor size would be given 84 Gy while a smaller tumor size would be given 119 Gy. At the second stage, in addition to the baseline risk factors, we also included the ALBI change from the baseline as a covariate. The estimated optimal dosage regime shows that, for patients with a lower planned dose of the first stage (<78 Gy), 55 Gy should be given. For patients with a higher planned dose (>78 Gy), the patients with larger ALBI change should be given a lower dose (25 Gy), and vice versa (44 Gy with smaller ALBI change).

Our recommendation that giving patients with a larger ALBI change a lower second stage dose is consistent with what has been reported in the literature, including that mid-treatment ALBI change in the analysis improved the ability to predict the liver toxicity (*Jackson et al.*, 2019) and greater mid-treatment increases in ALBI were associated with decreased overall survival (*Morris et al.*, 2019). ALBI is also recommended in practice since it is easily obtained from standard lab tests and is more sensitive than other commonly used biomarkers (*Mohammadi et al.*, 2018). Our results revealed some potential predictive biomarkers on which the optimal dosage regime should depend, such as AST and ALT at baseline and ALBI change at the second treatment stage. These results give radiologists a way to design randomized dose-response studies in the future to confirm the findings and to better treat patients.

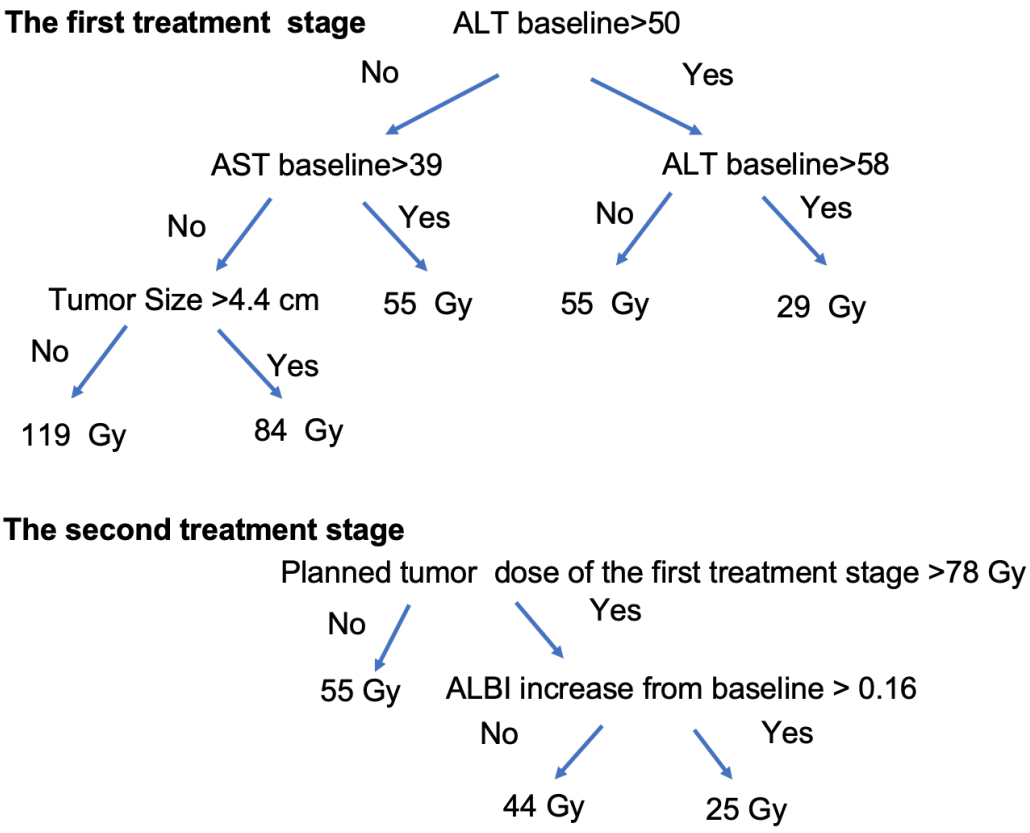

Figure 3.1: The estimated optimal dosage regime for patients with liver cancer from Michigan Medicine adaptive stereotactic body radiation therapy dataset

In addition to the consistent findings with clinical experience, the patients from this study would benefit from our estimated regimes by increasing their overall survival time when receiving the estimated optimal dosage regime. According to our model, the mean overall survival (in days) of the population would increase from 997 to 2190 days after utilizing the estimated optimal dosage regime at both treatment stages.

3.7 Discussion

To overcome the limitations of current methods when dealing with the dose-finding problem within the framework of a dynamic dosage regime, we developed the KIDD-Learning to estimate the personalized optimal dosage regime by accommodating pa-

tients' time-varying medical characteristics. The results from simulation studies have shown that KIDD-Learning has consistently superior performance compared to existing methods, in terms of maximizing the expect ed counterfactual outcomes and correctly assigning the optimal dosage regimes. Unlike other flexible dosage rules assignment methods (*Chen et al.*, 2016, 2018), the tree-based decision rules are easily interpreted and implemented. In particular, with the estimated decision rule, when a patient is newly presented at a clinic, a physician can easily determine the recommended dose for this patient based on his/her characteristics without additional calculation. Compared with LZ, a method also yields tree-based decision rules, KIDD-Learning is more robust under different scenarios when using observational data.

We also illustrate the potential application of KIDD-Learning to a Michigan Medicine liver cancer SBRT dataset. The dosage decision for each stage of a multiple-stage adaptive radiation therapy is currently more experience-based rather than based on evidence-based, quantitative tools, with no previous statistical methods to handle such a problem. KIDD-Learning uses a kernel-involved decision learning approach to estimate the optimal dosage strategy backwards from the second stage; therefore, the optimal dose assignment for the first stage is determined conditional on the already optimized outcome of the second stages. In this way, KIDD-Learning is an effective tool for physicians to provide optimal personalized radiation therapy over time.

There are a few directions for future work. One possible direction concerns the limitation of the dose-search algorithm we used to estimate the tree-based decision rule of each treatment stage. This tree-based dose-search algorithm, which splits the nodes from the top root node to terminal nodes, is based on a greedy algorithm, because the optimal decisions are made at each node without looking ahead to further child nodes. If there are strong predictors hidden from relatively weak imperceptible nodes, this top-down greedy algorithm may fail to find the optimal splitting rules

and fall into some sub-optimal rules. One way potentially to eliminate this greediness is "lookahead" (*Murthy and Salzberg*, 1995) by evaluating the future child nodes before splitting on their parent node. We tried "lookahead" in our preliminary simulation studies; the improvement of performance was moderate while the computational time increased substantially. To reduce the potential local optimality brought by the greedy top-down algorithm, some alternative search methods have been developed in computer science literature, such as stochastic tree search using the methods of Markov chain Monte Carlo (*Denison et al.*, 1998; *Chipman et al.*, 1998; *Wu et al.*, 2007; *Wang et al.*, 2017) and evolutionary algorithm (*Papagelis and Kalles*, 2001). In a future study, we will improve KIDD-Learning by incorporating these alternative search methods and providing more robust optimal decision rules.

Another possible exploration would address the potentially competing outcomes for determining the optimal dosage regime, such as efficacy, toxicity, side-effects and cost burden. In the data illustration of Michigan Medicine Liver SBRT dataset, we consider the overall survival as the outcome, as it is a combination of treatment effectiveness and toxicity. However, when a composite outcome, like overall survival, is not available, a method that is capable of balancing multiple contradictory objectives is desirable (*Butler et al.*, 2017). Some statistical methods have been developed to compromise between multiple contradictory outcomes (*Lizotte and Laber*, 2016; *Wang et al.*, 2019), but none is able to estimate the optimal dose on a continuum within the framework of dynamic dosage regime. One potential future direction is to further incorporate these multiple objective optimization functions and extend our KIDD-Learning.

CHAPTER IV

Stochastic Spline-Involved Tree Search for Optimizing Personalized Multi-stage Dosing Strategy

4.1 Introduction

Determining the optimal dosage is of paramount importance in drug development and personalized medicine (*Schmidt*, 1988; *Thall et al.*, 2007; *Thall*, 2008; *Guo and Yuan*, 2017; *Xu et al.*, 2018). An optimal treatment dose usually balances effectiveness and side-effects. Increasing the drug dose can intensify the treatment effect at the cost of damaging the function of normal organs, while a low drug dose may result in a lack of efficacy. Moreover, as the interest in precision medicine has emerged (*Collins and Varmus*, 2015), scientists are more interested in providing individualized doses for patients with heterogeneity, rather than giving a one-size-fits-all optimal treatment recommendation. Recently, statistical methods have been developed to accommodate the need for estimating the personalized optimal dose for each individual (*Li et al.*, 2019; *Guo and Yuan*, 2017; *Thall et al.*, 2008; *Xu et al.*, 2018; *Rich et al.*, 2016).

Moreover, when treating chronic diseases, such as cancer or diabetes, clinical decision making often involves the determination of overall dose level and the allocation of drug doses over several treatment stages. Doctors routinely provide treatments

based on patients' baseline covariates and time-varying characteristics. For example, in treating patients with diabetes, insulin dose titration has become a common practice to balance the desired sufficient insulin effectiveness and a patient's tolerance of vomiting and nausea (*Scarpello*, 2001; *Qu et al.*, 2019). Dose titration allows patients to start from a low insulin dose and gradually escalate to higher dose levels according to their tolerance to the drug. Such a treating process involves a sequence of dose assignments chosen based on up-to-date patient information. This kind of idea of personalizing treatment decisions over time has been formalized as the Dynamic Treatment Regimes (DTRs) (*Murphy et al.*, 2001; *Wang et al.*, 2012). DTRs generalize the individualized medical decisions into a time-varying treatment setting and thus accommodate the updated information for each person at each stage. In the remaining part of this article, we will refer to dynamic treatment regimes with continuous dose treatments as dynamic dosage regimes.

There has been a rapidly growing literature in estimating the optimal multiple-stage DTR with binary treatment or finite multiple treatments (*Zhang et al.*, 2015; *Laber and Zhao*, 2015; *Chen et al.*, 2018). However, it is complex to extend existing methods to personalized dose-finding because of the sparse nature of the observed data, i.e., the dose level follows a continuous distribution – of which the probability of observing any specific dose is zero. Estimating the optimal DTR for continuous treatment has received less attention in the literature; some examples include *Lee et al.* (2015); *Chen et al.* (2016, 2018); *Schulz and Moodie* (2020). However, many of these methods have limited interpretability, which hinders their potential impact. *Laber and Zhao* (2015) (denoted as LZ hereafter) proposed a tree-based method for estimating optimal DTR with continuous treatments, and demonstrated the promising performance under various scenarios. However, these methods require specifying the parametric working model for the treatment assignment mechanism, or the conditional outcome, or both. The quality of estimated DTRs is tied to how adequately

the working models approximate the true data-generating mechanism. In practice, it is difficult to correctly specify the outcome regression model even with background knowledge of the observed population.

In Chapter III, KIDD-Learning and implemented it using a modified tree-based reinforcement learning method. KIDD-Learning can estimate the optimal tree-based dosage regime. However, when implementing KIDD-Learning, the tree-based decision rules are constructed using a modified Classification and Regression Tree (CART) dose search algorithm. In CART, as the decision tree grows from the root node, a split is determined when the expected counterfactual outcome increases, and the splitting process stops after reaching the prespecified stopping criteria. However, the CART dose search algorithm is limited by its greedy nature, because splits are myopic and may fail to account for the possible impact of future partitions. Such a top-down algorithm generates a sequence of trees, each of which is a direct extension of the previous decision tree, and inevitably results in a locally optimal tree if the previous tree is already sub-optimal.

In addition, most of the existing methods listed above limit their application to data from randomized dose trials, where the confounders have been well-adjusted. Real world observational data are more common, because of the cost and the ethical concerns (*Wallace et al.*, 2018). Applying the methods designed for randomized trials directly without carefully adjusting for confounders may lead to a sub-optimal dosage regime. Specifically, in the method developed by *Chen et al.* (2016) (denoted as CZK hereafter), it was claimed that CZK remained valid by using an inverse weight of the density model conditional on the covariates from the observed data. However, their simulation results showed that the performance of CZK did not change much with or without adjusting for confounders. CZK's potential extension to using observational data thus needs further exploration and validation.

Therefore, a robust and interpretable method for estimating the optimal dynamic

dosage regime using observational data is highly desirable, as it bridges the gap between the clinician's medical expertise and data-driven dosage regimes, and allows a clinician to better understand and trust the regimes. In this article, we propose a stochastic spline-involved tree search learning method, SSITS for estimating the optimal dosage regime, which is an extension of KIDD-Learning. This new method combines a robust non-parametric estimation of the dose-response function with an efficient simulated annealing tree search algorithm for estimating interpretable dose decision rules over multiple decision stages. At each treatment stage, our proposed method stochastically searches the binary decision tree space and then provides the optimal decision rule by comparing the maximum value of each estimated dose-response function. Compared to its CART counterpart, SSITS can search the tree space more efficiently and is more capable of escaping the local optimality. In the cases of multiple-stage multiple-treatment DTR, the effectiveness of the stochastic search has been demonstrated by *Sun* (2019). Similar superiority is expected in SSITS when estimating the continuous dose DTR. In addition, unlike KIDD-Learning where the dose-response function has to be fitted to evaluate every possible splitting, this proposed method only fits the non-parametric regression model in the resulting terminal nodes for each visited tree and therefore is more computationally efficient than KIDD-Learning.

This article is organized as follows: Section 4.2 formalizes the continuous dose treatment DTR problem and connects it with observational data; Section 4.3 and 4.4 describes the proposed stochastic spline-involved tree search learning method for estimating the optimal dosage regime and outlines its implementation; Section 4.5 presents the numeric results from simulation studies; Section 4.6 illustrates the data application of SSITS using data from the International Warfarin Pharmacogenetics Consortium. Section 4.7 concludes the study with a discussion.

4.2 Data and Mathematical Formulation of Dynamic Dosage Regime

4.2.1 Notation and the Statistical Problem

We assume there are T stages in total. At each stage $t \in \{1, 2, \ldots, T\}$, we denote the continuous dose value of the drug taken at stage t by $D_t \in \mathcal{D}_t$, and the observed value of D_t is d_t. Without loss of generality, we further assume the domain of the dose assignment $\mathcal{D}_s = [0, 1]$, and $d_t \in \mathcal{D}_t$. Let $\mathbf{X}_t$ denote the patient information accumulated between stage $t-1$ and t, and the patient history up to stage t can be denoted as $\overline{\mathbf{X}}_t = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t)$. Similarly, the treatment history until stage t can be denoted as $\overline{\mathbf{D}}_t = (D_1, \ldots, D_{t-1})$. The outcome of interest is denoted by Y. In practice, the reward outcome we are interested in might be a combination of some intermediate outcomes, or the outcomes happen after some intermediate stages may have an impact on the final outcome. Thus, the overall outcomes vector is $(Y_1, Y_2, \ldots, Y_T)$ and the outcome history $\overline{\mathbf{Y}}_{t-1} = (Y_1, \ldots, Y_{t-1})$. We further denote that the final outcome of interest Y is a function of all intermediate outcomes, i.e., $Y = s(Y_1, Y_2, \ldots, Y_T)$, where $s(\cdot)$ is a pre-specified function (e.g., sum). We assume Y to be bounded and higher value of Y is more desirable. We define the history prior to the dose treatment at stage t D_t as $\mathbf{H}_t$. which includes the dose assignments $\overline{\mathbf{D}}_{t-1}$ history, the history happened before stage t, $\overline{\mathbf{X}}_t$, and the intermediate outcomes. Specifically, $\mathbf{H}_t = \{(\overline{\mathbf{D}}_{t-1}, \overline{\mathbf{X}}_t, \overline{\mathbf{Y}}_{t-1})\}_{i=1}^n$. Therefore, across patient index i, the observed data consist n i.i.d. trajectories are denoted as $\{D_{t,i}, \mathbf{H}_{t,i}, Y_{t,i}\}_{t=1}^T$. For brevity, we suppress the subject index i in the following text when no confusion exists.

The goal of our method is to use the observed data to find the optimal dosage regime that determines, based on a patient's medical history, what dose he/she should receive at each stage. We formally define a personalized decision rule sequence $\mathbf{g} = (g_1, g_2, \ldots, g_T)$ as the dynamic dosage regime, where g_t, the decision rule of

stage t, maps the observed history $\mathbf{H}_t$ about the patient's health history to the dose assignment, i.e.,

$$\mathbf{H}_t \mapsto g_t(\mathbf{H}_t) \in \mathcal{D}_t,\ t \in \{1, 2, \ldots, T\}$$

The counterfactual framework of causal inference proposed by *Robins* (1986) is followed to identify the optimal dynamic dosage regime from the observed data. The counterfactual outcome if a patient follows the dynamic treatment regime $\mathbf{g}$ is denoted by $Y^*(\mathbf{g})$. $E(Y^*(\mathbf{g}))$ is the expectation with respect to the distribution of patients' history of the counterfactual outcome if the entire population had followed the given dosage regime $\mathbf{g}$. The optimal dosage regime $\mathbf{g}^{opt}$ should be a sequence of decision rules that is able to make the entire population benefit from it. Therefore, we formally define the optimal $\mathbf{g}^{opt}$ as a sequence of decision rule that can lead to the optimal value of $E\{Y^*(\mathbf{g})\}$, such that,

$$E\{Y^*(\mathbf{g}^{opt})\} \geq E\{Y^*(\mathbf{g})\},\ \forall\, \mathbf{g} \in \mathcal{G}, \tag{4.1}$$

where $\mathcal{G}$ is the set of all potential dosage regimes under consideration.

4.2.2 How to Use the Observational Data

In Eqn (4.1), the expectation of counterfactual is utilized to estimate the optimal dosage regime. Considering the limited accessibility of counterfactual outcome, which often comes from the data of randomized clinical trials, we have to connect the counterfactual outcome with the data we observe by applying the framework and assumptions of causal inference (*Robins*, 1986). In particular, the method to estimate the whole optimal dynamic dosage regime includes backward induction; therefore, we start formulating this optimization problem from the final stage T in a reverse sequential order. At the final stage T, let $Y^*(D_1, D_2, \ldots, D_{T-1}, d_T)$, or $Y^*(d_T)$ for brevity, denote the counterfactual outcome if a patient takes dose d_T conditional on

his/her previous medical characteristics. In order to better connect the counterfactual outcome $Y^*(d_T)$ with the observed data $\{D_T, \mathbf{H}_T, Y\}$, we make the following causal inference assumptions:

Assumption 1 *Consistency*: $D_T = d_T$ implies $Y = Y^*(d_T)$. This assumption ensures the observed outcome is the same as the counterfactual outcome under the dose actually assigned.

Assumption 2 *Exchangeability*: $Y^*(d_T) \perp D_T \mid \mathbf{H}_T, \forall\, d_T \in \mathcal{D}_T$, where $\perp$ denotes statistical independence. This assumption guarantees that the counterfactual outcome under a certain dose d_T is independent of the dose choice, conditional on the previous history $\mathbf{H}_T$.

Assumption 3 *Positivity*: $\pi(d_T \mid \mathbf{H}_T) \geq \pi_{\min} > 0$ for all possible $\mathbf{H}_T$, where $\pi(d_T \mid \mathbf{H}_T) = \frac{\partial P(D_T \leq d_T | \mathbf{H}_T)}{\partial d_T}$ is the conditional treatment density given history $\mathbf{H}_T$ and $\pi_{\min}$ is a given positive number. It ensures that every patient has some chance of receiving a certain treatment d_T

Under the assumptions above, the expectation of counterfactual outcome $Y^*(d_T)$ can be identified from the observed data as

$$E\{Y^*(d_T)\} = E\{E(Y \mid D_T = d_T, \mathbf{H}_T)\} = \int_{\mathcal{H}_T} E(Y \mid D_T = d_T, \mathbf{H}_T = \mathbf{h}_T)\, \mathrm{d}P(\mathbf{h}_T).$$

Dose-response function is a natural way to visualize the dose-response relationship and researching the continuous dose effect (*Kennedy et al.*, 2017; *Zhao et al.*, 2020), therefore, we denote $\theta_T(d_T) = E(Y^*(d_T))$, and $\theta_T(\cdot)$ is the dose-response function of the population we are interested in. To guarantee that the optimal dose is identifiable within the domain, we have another assumption:

Assumption 4 *Concavity*: $\theta_T(\cdot)$ is continuous on the closed interval of $\mathcal{D}_T$ and differentiable on the open interval of $\mathcal{D}_T$; the second derivative $\theta''_T(\cdot)$ exists

throughout $\mathcal{D}_T$ and $\theta_T''(d_T) < 0 \; \forall \, d_T \in \mathcal{D}_T$. This assumption ensures the existence of the local maximum of the dose-response function within the domain $\mathcal{D}_T$.

Let $\theta_T\{g_T(\mathbf{H}_T)\}$ denote the dose response of the outcome under regime g_T. According to Eqn (4.1), the optimal regime of the final stage $g_T^{opt}(\mathbf{H}_T)$ satisfies

$$E[Y^*\{g_T^{opt}(\mathbf{H}_T)\}] \geq E[Y^*\{g_T(\mathbf{H}_T)\}], \; \forall \, g_T \in \mathcal{G}_T.$$

i.e.

$$\begin{aligned} g_T^{opt}(\mathbf{H}_T) &= \arg\max_{g_T \in \mathcal{G}_T} E\{E(Y \mid D_T = g_T(\mathbf{H}_T), \mathbf{H}_T)\} \\ &= \arg\max_{g_T \in \mathcal{G}_T} \int_{\mathcal{H}_T} E(Y \mid D_T = g_T(\mathbf{H}_T), \mathbf{H}_T = \mathbf{h}_T) \mathrm{d}P(\mathbf{h}_T) \\ &= \arg\max_{g_T \in \mathcal{G}_T} \theta_T\{g_T(\mathbf{H}_T)\}. \end{aligned} \tag{4.2}$$

At the intermediate stage $t \in \{1, 2, \ldots, T-1\}$, it is important to compare the decision rules of a intermediate stage when the decision rules after the current stage are the same, unless the comparison becomes meaningless. Thus, we compare the decision rule of the current stage while assuming their future dose assignments have been optimized. We denote the future optimized counterfactual outcome with all future stages dose assignments under the optimal decision rules, when a patient at stage t receiving treatment d_t as $\{Y^*(D_1, \ldots, D_t, g_{t+1}^{opt}, \ldots, g_T^{opt})_{D_t = d_t}\}$, for brevity $\widetilde{Y}_t^*(d_t)$.

Under similar causal inference assumptions, the expectation of the future optimized counterfactual outcome can be identified through observed data as $E(\widetilde{Y}_t^*(d_t)) = E\{E(PO_t \mid D_t = d_t, \mathbf{H}_t)\}$, where PO_t denotes pseudo-outcome of stage t. Specifically, we define PO_t recursively using Bellman's optimality as $PO_t = \hat{E}\{PO_{t+1} \mid D_{t+1} = g_{t+1}^{opt}(\mathbf{H}_{t+1}), \mathbf{H}_{t+1}\}$ and $PO_T = Y$.

Similarly, we can define at any stage $\theta_t(d_t) = E\{\widetilde{Y}_t^*(d_t)\}$, where $\theta_t(\cdot)$ is the dose-response function of the entire population at stage t, and the concavity of $\theta_t(\cdot)$ should

be assumed throughout the domain of $\mathcal{D}_t$. Thus, the dose-response function at stage t under a decision rule $g_t(\mathbf{H}_t)$ is denoted as $\theta_t\{g_t(\mathbf{H}_t)\}$. $g_t^{opt}(\mathbf{H}_t)$ should satisfy $E[\widetilde{Y}_t^*\{g_t^{opt}(\mathbf{H}_t)\}] \geq E[\widetilde{Y}_t^*\{g_t(\mathbf{H}_t)\}]$, $\forall\, g_t \in \mathcal{G}_t$, we have the optimization problem at the intermediate stage t becoming

$$\begin{aligned} g_t^{opt}(\mathbf{H}_t) &= \arg\max_{g_t \in \mathcal{G}_t} \int_{\mathcal{H}_t} E\{PO_t \mid D_T = g_t(\mathbf{H}_t), \mathbf{H}_t = \mathbf{h}_t\} \mathrm{d}P(\mathbf{h}_t) \\ &= \arg\max_{g_t \in \mathcal{G}_t} \theta_t\{g_t(\mathbf{H}_t)\}. \end{aligned} \tag{4.3}$$

4.3 Stochastic Spline-Involved Tree Search for Optimizing Personalized Multi-stage Dosage Regime

Decision tree is one of the most popular models for interpretable statistical learning. To estimate a tree-based decision rule, the tree-based reinforcement learning method (*Laber and Zhao*, 2015; *Tao et al.*, 2018) can be a natural option. LZ recursively splits the patient feature space by maximizing the purity measure at each split. Such a top-down, greedy algorithm may end up estimating a local optimal decision rule when there are strong predictors hidden from imperceptible parent nodes. One way to alleviate the greediness is lookahead by finding new splits based on optimizing deeper trees rooted at the current terminal node, rather than just optimizing a single split (*Murthy and Salzberg*, 1995; *Tao et al.*, 2018). However, lookahead is still heavily limited by its local move, since it generates a sequence of trees, each of which is an extension of the previous tree. It does not resolve the problem of local optimality while increasing the computational burden substantially (*Bertsimas and Dunn*, 2017; *Tao et al.*, 2018).

Constructing an optimal decision tree is known to be NP-complete (*Laurent and Rivest*, 1976), here we propose to use stochastic tree search to expand the search scope and to better balance the exploration and the exploitation. Within each terminal node, a flexible smooth coefficient model is used to model the dose-response function

while adjusting for confounders.

4.3.1 Estimating the Dose-response Function

The estimator of the dose-response is vital in assessing the quality of estimated dosage strategy. To estimate the dose-response function using observational study data, the key is to adjust for the confounding variables. *Rosenbaum* (1987) developed the propensity score method to address the confounding bias for binary treatment assignment, and other statisticians have proposed generalizations of the propensity score method to deal with continuous treatment (*Imai and Van Dyk*, 2004; *Hirano and Imbens*, 2005; *Flores et al.*, 2012). In particular, the propensity function (PF) used a generalized propensity score-based method for causal inference with continuous treatments (*Imai and Van Dyk*, 2004).

The propensity function $\pi_\psi(D_t \mid \mathbf{H}_t)$ is defined as the conditional density function of a dose assignment given the observed history at stage $t \in \{1, \ldots, T\}$, where ψ parameterizes this distribution. We further fit the parametric model of the propensity function as $e_\psi(\cdot \mid \mathbf{H}_t) = \pi_\psi(\cdot \mid \mathbf{H}_t)$. In addition to the assumptions from Section 2 above, we make another assumption to simplify the representation of PF. We assume a *Uniquely Parameterized Propensity Function.* In particular, for every value of $\mathbf{H}_t \in \mathcal{H}_\mathbf{t}$, there exists a unique finite-dimensional parameter $\boldsymbol{\phi} \in \Phi$, such that $e_\psi(\cdot \mid \mathbf{H}_t)$ depends on $\mathbf{H}_t$ only through $\boldsymbol{\phi}_\psi(\mathbf{H}_t)$. In other words, the propensity function $e(\cdot \mid \boldsymbol{\phi}_\psi(\mathbf{H}_t))$ can be uniquely represented by $\boldsymbol{\phi}$. Therefore, we can rewrite PF as $e(\cdot \mid \boldsymbol{\phi})$, where $\boldsymbol{\phi} = \boldsymbol{\phi}_\psi(\mathbf{H}_t)$. This parametrization assumption that uses $\boldsymbol{\phi}$ to represent $\mathbf{H}_t$ can decrease the dimension in future calculations when adjusting for confounders, as $\boldsymbol{\phi}$ has typically much lower dimension than $\mathbf{H}_t$. Without loss of generality, we specify $\boldsymbol{\phi}$ as univariate, and rewrite $\boldsymbol{\phi}$ as ϕ. In the case of continuous treatment, we can assume the treatment $D_t \sim N(\beta \mathbf{H}_t, \sigma^2)$, where $\psi = (\beta, \sigma^2)$. Finally, $\phi_\psi(\mathbf{H}_t) = \beta \mathbf{H}_t$ can uniquely represent the propensity function $e(\cdot \mid \phi_\psi(\mathbf{H}_t))$.

4.3.1.1 Use a smooth coefficient model to estimate the dose-response function

The dose response can be estimated by matching or subclassifying on the propensity function. However, the method we may use in subclass models may be constrained due to the dependency on the parametric assumptions within the subclass models; therefore, to better understand the dose-response relationship in practice, we prefer estimating a robust and flexible dose-response function. We further fit the dose-response model using a smooth coefficient model (SCM), also known as varying coefficient model (*Hastie and Tibshirani*, 1993; *Zhao et al.*, 2020). This model allows the regression coefficients to vary smoothly and systematically in more than one dimension as a function of the propensity function and the dose assignment D_t.

At the final stage T, according to the uniquely parameterized assumption of the propensity function, the dose-response function can be re-written as

$$\theta_T(d_T) = E(Y^*(d)) = E\{E(Y \mid D_T = d_T, \mathbf{H}_T)\} = E\{E(Y \mid D_T = d_T, \phi_T)\}, \quad (4.4)$$

where $\phi_T = \phi_\psi(\mathbf{H}_T)$ is the propensity function at the final stage T. We can model the rightmost part of Eqn 4.4, $E(Y \mid D_T = d_T, \phi_T)$ using a smooth coefficient model, i.e.,

$$E(Y \mid D_T = d_T, \phi_T) = f(\phi_T, D_T),$$

where $f(\cdot)$ is a smooth function of ϕ_T and D_T. In practice, ϕ_T is replaced by $\widehat{\phi}_T$ from the fitted propensity function. Therefore, we can average over the empirical distribution of ϕ_T to obtain an estimate of the dose-response function using a SCM of the PF, such as

$$\widehat{\theta}_T(d_T) = \widehat{E}[Y^*(d_T)] = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}(\widehat{\phi}_{T,i}, d_T), \quad (4.5)$$

where $\widehat{f}(\cdot)$ is the fitted SCM and $\widehat{\phi}_{T,i}$ is from the fitted propensity function. The dose-response function can be evaluated in Eqn 4.5 separately, using a pre-specified grid of dose values within the domain $\mathcal{D}_T$.

At any intermediate stage $t \in \{1, 2, \ldots, T-1\}$, the future optimized $Y_t^*(d_t)$ is not observable and needs to be replaced with the future optimized pseudo outcome as defined previously. The estimation of $\widehat{\theta}_T(d_T)$ above works for any stage $t \in \{1, 2, \ldots, T-1\}$ prior to the final stage T. In the implementation, the SCMs are fitted using the R package "*mgcv*" (*Wood*, 2017), and the smooth functions are represented as a weighted sum of known basis functions, which are penalized cubic regression splines with the number of knots equal to 5. Cubic spline is utilized as it is flexible but also smooth enough to capture the curve. According to our preliminary simulation studies, we varied the number of knots from 5 to 10, and the estimation of the dose-response function stayed robust.

4.3.2 Simulated Annealing Algorithm for Stochastically Searching the Optimal Dosage Regime

Simulated annealing is an algorithm inspired by a powerful optimization process that happens during the cooling of physical systems (*Salter and Pearl*, 2001; *Wang et al.*, 2017). The goal of the cooling process is to obtain a solid that has minimal energy. This intuition can be generalized to other optimization goals; in our case, the goal is to find an optimal dosage regime that maximizes the expected counterfactual dose response outcome of the whole population. In practice, the optimal decision rule of each treatment stage within the dosage regime is estimated backwards in a reverse sequential order; therefore, in the following derivation, the algorithm works for any treatment stage t; for convenience, we suppress the stage subscript t when no confusion exists. Formally, we denote a tree-based continuous dose decision rule as g. g can be evaluated using a value function $F(g)$, which is intended to be optimized

to identify the optimal decision rule. The simulated annealing algorithm proceeds by walking stochastically through the tree space in such a way that a decision rule g that increases the value of $F(g)$ is always accepted while those that decrease the value can be accepted with a specific probability. This probability depends on the number of iterations and other pre-specified hyper parameters. The simulated annealing algorithm may accept a worse solution in order to help the optimization process proceed and avoid being trapped in a local optimality.

Let $\mathbf{T}_0$ be the initial temperature parameter, and let N_{iter} be the total number of iterations. These two hyper parameters and the formula of temperature change control the rate of identifying the optimal solution and are critical for obtaining a robust optimal solution. In particular, if the temperature decreases too quickly, the algorithm might be trapped within a local optimal point, while the slower temperature decrease can require excessive computational efforts. In our implementation, we set $\mathbf{T}_0 = 100$. Let m be a specific iteration, $m \in 1, 2, \ldots, N_{iter}$, let g^m be the tree-based decision rule at iteration m, and let $\mathbf{T}^m$ be the temperature at iteration m. The following steps are repeated until m reaches N_{iter}:

1. For the current dose decision rule g^m, generate a potential decision rule $g^{m\prime}$ from g^m according to the generation procedure;
2. Calculate $F(g^{m\prime})$;
3. Accept $g^{m+1} = g^{m\prime}$ with probability $\min\{1, \exp(\frac{F(g^{m\prime})-F(g^m)}{T^m})\}$, otherwise $g^{m+1} = g^m$;
4. Update the temperature $\mathbf{T}^m = \mathbf{T}_0^{1-\frac{m}{N_{iter}}}$.

We further specifically denote the dose decision rule g as $(\mathcal{P}, \mathcal{R})$, where $\mathcal{P}$ is the parameter set that describes a tree topology, and $\mathcal{R}$ is the optimal dose assignment rule for each terminal node of $\mathcal{P}$. At each iteration m, the new solution can be

generated via two steps: first generate a tree topology $\mathcal{P}^m$, and then identify the dose assignment rule $\mathcal{R}^m$ based on $\mathcal{P}^m$. Thus, g^m can be evaluated and the acceptance can be determined accordingly.

When generating $g^{m\prime}$ from the current solution g^m, we have to first generate the tree topology $\mathcal{P}^{m\prime}$ from $\mathcal{P}^m$. In particular, a tree topology consists of three elements: tree arrangement, splitting variables and splitting thresholds. Our algorithm proposes generating $\mathcal{P}^{m\prime}$ from a solution that is neighboring to $\mathcal{P}^m$, rather than generating it uniformly. To be more specific, the neighboring tree topology we use is generated by five different proposals, namely, GROW, PRUNE, CHANGE, SWAP and RESTRUCTURE (*Denison et al.*, 1998; *Chipman et al.*, 1998; *Wu et al.*, 2007). The GROW proposal randomly picks one terminal node from the current tree $\mathcal{P}$ and splits it into two new terminal nodes. The PRUNE proposal is the reverse of GROW as it randomly picks one parent node with two terminal nodes and collapses its terminal nodes and turns the parent node into a terminal node. The CHANGE proposal randomly picks a non-terminal internal node and resamples the splitting rule. The SWAP proposal randomly picks a pair of parent-child internal nodes and swaps their splitting rules. Considering the restricted change these proposals make to the current $\mathcal{P}^m$, we also propose a more radical change of the tree structure, called RESTRUCTURE (*Wu et al.*, 2007). The RESTRUCTURE proposal modifies the structure of the internal nodes while maintaining the number of terminal nodes and the partition of the subjects. This radical change helps the proposed decision tree escape from the local optimality and search more efficiently within the tree topology space. The five proposals are implemented with pre-specified probabilities.

For a given tree topology $\mathcal{P}$, the next step is to identify the optimal dose assignment rule $\mathcal{R}$. Let $b(\mathcal{P})$ denote the set of all terminal nodes and $v \in b(\mathcal{P})$; $\mathcal{R}$ denote the dose assignment rule that assigns specific dose d_v to each terminal node v. We can further define the objective function $F(g)$ as a sum of each subpopulation dose

response, such as, $F(g) = \sum_{v \in b(\mathcal{P})} \hat{\theta}_v(d_v)\mathbb{P}_n[I(\mathbf{H} \in v)]$, where $\mathbb{P}_n$ is the empirical expectation operator, and $\hat{\theta}_v(d_v)$ is the estimated expected counterfactual dose response if all subjects in terminal node v take dose d_v. Since the identification of the optimal decision rule happens by maximizing the expected counterfactual dose response, assigning any random dose d_v to each terminal node v is reasonable but not efficient. In this case, a large proportion of the population does not receive their optimal dose, and thus the random dose assignment does not maximize our objective in the given tree topology. Alternatively, in order to maximize $F(g)$, $\mathcal{R}$ assigns dose d_v^{opt} to each terminal node v, where $d_v^{opt} = \arg\max_{d_v \in \mathcal{D}} \theta_v(d_v)$. Thus, $F(g)$ can be re-written as

$$F(g) = \sum_{v \in b(\mathcal{P})} \hat{\theta}_v(d_v^{opt})\mathbb{P}_n[I(\mathbf{H} \in v)]. \tag{4.6}$$

Since the concavity of $\theta_v(\cdot)$ ensures the existence of the maximum of the dose-response function, the treatment assignment of each terminal node v can be uniquely determined. Thus the optimal tree-based decision rule can be identified from the stochastically generated rule sequence via the simulated annealing algorithm. The details are summarized as in Algorithm 4 below.

4.4 Implementation of SSITS

The proposed method is implemented by backward induction. At each stage, the optimal dose decision is determined by stochastically evaluating the expected counterfactual dose response outcome of the proposed tree-based decision rule. Before estimating the dose-response function of each stage t, we have to estimate the pseudo outcome PO_t first, which is the replacement of Y in the intermediate stage. At the final stage $t = T$, $PO_t = Y$, and it can be used directly to estimate the dose-response function, thus to estimate the regime. At any intermediate stage $t \in \{1, 2, \ldots, T-1\}$, the pseudo-outcome PO_t depends on an estimated dosage regimes where the dose

assignment of the future stages have been already optimized, as $PO_t = \hat{E}\{PO_{t+1} \mid D_{t+1} = g_{t+1}^{opt}(\mathbf{H}_{t+1}), \mathbf{H}_{t+1}\}$. We further assume the dose effect Y_t is cumulatively carried forward to the final outcome Y. To reduce the accumulated bias, we calculate the pseudo outcome PO_t by using Y_t, the actual observed intermediate outcome at stage t plus the difference of the expected optimal dose-response pseudo outcome and the expected pseudo outcome of actual observed of the stages after t; that is

$$PO_t = Y + \sum_{j=t+1}^{T} \{E(PO_j \mid \mathbf{H}_j, D_j = g_j^{opt}(\mathbf{H}_j)) - E(PO_j \mid \mathbf{H}_j, D_j = d_j)\},$$

where $PO_T = Y$ and PO_t is calculated by using off-shelf machine-learning method "*eXtreme Gradient Boosting*" (*Chen and Guestrin*, 2016). Note here, the intermediate outcomes Y_t might be different measures of patient's disease status, such as toxicity or efficacy, because of the different priority of assigning a dose during the progression of disease. Therefore, before aggregating various intermediate outcomes from different stages, the scale should be standardized by clinician's suggestion. The implementation of Stochastic Spline-Involve Tree Search for estimating the optimal dosage regime with T treatment stage is outlined in Algorithm 4. Note here, N_{iter} is the pre-specified total number of simulated annealing iterations. In practice, to yield a more robust estimated dosage regime, the N_{iter} iterations can be repeated in multiple chains using different initial $\mathcal{P}^1$. To further promote randomness, in each chain we start from a random tree sampled from tree-generating process $\pi(\mathcal{P})$. Details on how to generate the initial tree topology can be found in the *Appendix*.

4.5 Simulation Studies

We conduct a number of simulation studies to assess the performance of SSITS. To facilitate the comparison with existing methods, we conduct simulation scenarios, and estimate the optimal dosage regime using methods proposed by CZK (*Chen et al.*,

Algorithm 4 Stochastic Tree Search for the optimal dosing in T-stage decision rules

Result: $\widehat{\mathbf{g}}^{opt} = \widehat{g}_1^{opt}, \ldots, \widehat{g}_T^{opt}$

Set $t = T$, $PO_T = Y$.

while $t \geq 1$ **do**

Set $m = 1$

Initialize $\mathcal{P}^1 \sim \pi(\mathcal{P})$ for stage t

while $m \leq N_{iter}$ **do**

1. Propose $\mathcal{P}^{m\prime}$ from $\mathcal{P}^m$;
2. Estimate the dose-response functions $\widehat{\theta}_{t,v}(d_v)$ using PO_t for $v \in b(\mathcal{P}^{m\prime})$;
3. Calculate $F(g^{m\prime}) = \sum_{v \in b(\mathcal{P}^{m\prime})} \hat{\theta}_{t,v}(d_v^{opt}) \mathbb{P}_n[I(\mathbf{H}_t \in v)]$;
4. Accept $g^{m+1} = g^{m\prime}$ with probability $\min\{1, \exp(\frac{F(g^{m\prime}) - F(g^m)}{T^m})\}$, otherwise $g^{m+1} = g^m$;
5. Update the temperature $\mathbf{T}^{m+1} = \mathbf{T}_0^{1 - \frac{m}{N_{iter}}}$;

$m = m + 1$.

end while

Estimate the optimal decision rule using $\widehat{g}_t^{opt} = \arg\max_{g_i : i \in \{1,2,\ldots,m\}} F(g_i)$

$t = t - 1$

Set $PO_t = Y + \sum_{j=t+1}^{T} \{E(PO_j \mid \mathbf{H}_j, D_j = g_j^{opt}(\mathbf{H}_j)) - E(PO_j \mid \mathbf{H}_j, D_j = d_j)\}$,

end while

2016) and LZ (*Laber and Zhao*, 2015). CZK is derived from outcome weighted learning (*Zhao et al.*, 2012) using non-linear kernels. It uses a non-convex loss function for the dose assignment optimization and solves the optimization problem by using the difference of convex functions. To accommodate the observational data, we use support vector regression, as presented by *Chen et al.* (2016), to estimate the conditional density of having one certain dose treatment. LZ is a tree-based reinforcement learning method that uses CART to estimate the optimal dosage regime. The purity measure of LZ is calculated by the kernel estimation using a plug-in bandwidth. The comparison with CZK and LZ is implemented using R packages "*kernlab*" (*Karatzoglou et al.*, 2004), "*SVMW*" (*Chen et al.*, 2016), "*gbm*" (*Ridgeway and Ridgeway*, 2004) and "*np*" (*Hayfield and Racine*, 2008). A variety of single- and two-stage scenarios are considered. For each scenario, we apply each method on a training set with a sample size N=300 or 500, and evaluate the estimated regimes on a separate test set with a sample size N=1000.

4.5.1 Single-Stage Scenarios

We consider single-stage scenarios with a continuous dose and generate baseline covariates $X_1, \cdots, X_k$ according to $uniform(0,1)$. k is the number of baseline variables. In practice, when the tailoring variables are known by investigators, k can be small. When the tailoring variables are not fully understood, k will be relatively large. The dose assignment D_1 is set within the range of $\mathcal{D}_1 = [0,1]$, generated from $Beta(\alpha, 1-\alpha)$ with $\alpha = 1/\{1+\exp(0.3X_1+0.2X_2+0.1X_3)\}$, which makes D_1 depend on the observed covariates.

To study the impact of underlying model specification on the performance of SSITS, two forms of the underlying optimal dosage regimes $g_1^{opt}(\mathbf{H}_1)$ are considered, one a tree-type dosage regime (Scenario 1, 2, 3), the other a non-tree-type regime (Scenario 4, 5, 6 and 7). The underlying tree-type is consistent with the dose as-

signment model when using the proposed stochastic tree search algorithm, while the models with underlying non-tree-type are "mis-specified." The detailed specification of $g_1^{opt}(\mathbf{H}_1)$ can be found in Table 4.1

These various non-tree-type scenarios can evaluate the robustness of SSITS against model mis-specification. In particular, the setting of Scenario 5 is inspired by *Chen et al.* (2016); the relationship between observed covariates and the optimal dose is non-linear and far away from a tree-type regime.

Our continuous outcome Y_1 has the form of $Y_1 = m(D_1) + \beta\mathbf{H_1} + \epsilon_1$, where $\beta\mathbf{H_1} = a\sum_{i=1}^{k} X_i + b$ and ϵ_1 is an independent standard normal variate that follows $N(0,1)$; $m(d_1) = 1.1 \times F_{24,6.75}[0.8 - (g_1^{opt} - d_1)]$, and $F_{p,q}(d_1)$ is a unimodal function within the domain to ensure the existence of the maximum dose effect. Specifically, $F_{p,q}(d_1) = (d_1)_1^{p-1}(1-d_1)^{q-1}\Gamma(p+q)/\{\Gamma(p)\Gamma(q)\}$ is the probability density function of $Beta(p,q)$.

In the data generation process, when the underlying true optimal dose is already known, we use the training dataset to estimate the optimal regimes and predict the optimal dose in the testing dataset. We use $\mid d^{opt} - \widehat{d}^{opt} \mid$ and $\widehat{E}(Y^*(\widehat{g}^{opt}))$ to evaluate the performance of the methods, where d^{opt} is the known true optimal dose under the pre-specified underlying optimal regime g^{opt}, and $\widehat{d}^{opt}$ is the estimated optimal dose calculated from the estimated optimal dosage regime $\widehat{g}^{opt}$. $\mid d^{opt} - \widehat{d}^{opt} \mid$ shows the average of how close the estimated optimal dose is to the true optimal dose; smaller is better. $\widehat{E}(Y^*(\widehat{g}^{opt}))$ is the expectation of the estimated mean counterfactual outcome, which indicates how much the whole population would benefit from the estimated optimal dose regimes if all subjects were to receive the estimated optimal dose. Since the results of $\widehat{E}(Y^*(\widehat{g}^{opt}))$ and $\mid d^{opt} - \widehat{d}^{opt} \mid$ are not normally distributed *Sun* (2019); *Tao et al.* (2018), we present the results of medians with the interquartile range from 25^{th} quartile to 75^{th} quartile.

Table 4.2 shows the single-stage performance of SSITS across the different sce-

Table 4.1: Specification of $g_1^{opt}(\mathbf{H}_1)$ for single stage simulation studies

Scenario	Type	Specification of $g_1^{opt}(\mathbf{H}_1)$
Tree Type	1	$0.8I(X_1 > 0.7) + 0.5I(X_1 \leq 0.7)I(X_2 > 0.5) + 0.2I(X_1 \leq 0.7)I(X_2 \leq 0.5)$
	2	$0.8I(X_1 > 0.8) + 0.4I(X_1 \leq 0.8)I(X_2 \leq 0.3)$ $+0.2I(X_1 \leq 0.8)I(X_2 > 0.3) + 0.6I(X_1 \leq 0.8)I(X_2 > 0.3)I(X_4 > 0.46).$
	3	$0.2I(X_1 \leq 0.55)I(X_4 \leq 0.46) + 0.4I(X_1 \leq 0.55)I(X_4 > 0.46)$ $+0.6I(X_1 > 0.55)I(X_3 > 0.6) + 0.8I(X_1 > 0.55)I(X_3 \leq 0.8)$
Non Tree type	4	$0.8I(X_1 + X_2 > 1.2) + 0.5I(X_1 + X_2 \leq 1.2)I(X_2 > 0.4)$ $+0.2I(X_1 + X_2 \leq 1.2)I(X_2 \leq 0.4)$
	5	$0.3I(X_1 < 0.5) + 0.6I(X_1 \geq 0.5) + X_2/3 + 0.25\log(\mid X_3 \mid +1) - 0.25$
	6	$0.8I(X_2/3 + 0.25\log(\mid X_3 \mid +1) > 0.25)I(\exp(x_1)^2 + X_5 \leq 3)$ $+0.6I(X_2/3 + 0.25\log(\mid X_3 \mid +1) > 0.25)I(\exp(x_1)^2 + X_5 > 3)$ $+0.2(X_2/3 + 0.25\log(\mid X_3 \mid +1) \leq 0.25)I(\exp(X_3)^2 + \log(X_4) \leq 0.66)$ $+0.4(X_2/3 + 0.25\log(\mid X_3 \mid +1) \leq 0.25)I(\exp(X_3)^2 + \log(X_4) > 0.66)$
	7	$0.9I(X_2/3 + 0.25\log(\mid X_3 \mid +1) > 0.3)I(\exp(x_1)^2 + X_5 \leq 3.2)$ $+0.6I(X_2/3 + 0.25\log(\mid X_3 \mid +1) > 0.3)I(\exp(x_1)^2 + X_5 > 3.2)$ $+0.2(X_2/3 + 0.25\log(\mid X_3 \mid +1) \leq 0.3)I(\exp(X_3)^2 + \log(x_4) \leq 0.56)$ $+0.5(X_2/3 + 0.25\log(\mid X_3 \mid +1) \leq 0.3)I(\exp(X_3)^2 + \log(X_4) > 0.56)I(X_2^2 + X_5^3 > 0.3)$ $+0.4(X_2/3 + 0.25\log(\mid X_3 \mid +1) \leq 0.3)I(\exp(X_3)^2 + \log(X_4) > 0.56)I(X_2^2 + X_5^3 \leq 0.3)$

narios described above. In particular, we set $k = 5$ to mimic the scenarios where the tailoring variables are clear to clinicians. Thus, a=1.41 and b=0.46. When the underlying dosage regime is correctly specified, i.e., tree-type, SSITS has consistently superior performance than LZ and CZK in all settings. Specifically, when sample size N=500, SSITS correctly assigns most of the patients into their optimal dose, as $\mid d^{opt} - \widehat{d}^{opt} \mid$ is close to 0 in Scenario 1, 2, and 3, respectively. The medians of the estimated mean counterfactual outcome of these three scenarios are also close to the pre-specified true value 10. When the sample size decreases to N=300, the performance gets worse but still outperforms LZ and CZK. The compromised performance of CZK is to be expected, as this method makes better use of the information of patients whose observed dose is close to the predicted optimal dose. In other words, if received dose is not close to the optimal dose for most subjects in the observed data, which is common in an observational study and is also the case in our simulation setting, CZK's performance is substantially undermined. In addition, the worse performance of LZ might be due to its fragility with severe model misspecification. LZ requires a correct specification of the conditional density of the model of treatment assignment, and also a correct specification of an outcome regression model. Mis-specification of these models may result in estimating a sub-optimal dosage regime.

When the underlying dosage regimes are mis-specified (Scenario 4, 5, 6, and 7), the performance of SSITS is undermined, but is still better compared to other existing methods. Furthermore, In scenario 5, 6 and 7 when SSITS largely outperform LZ and CZK in $E\{Y^*(\widehat{g}^{opt})\}$, the values in $\mid d^{opt} - \widehat{d}^{opt} \mid$ are close to their counterparts. The difference of performance is because SSITS identifies the optimal dosage regime by maximizing the expected counterfactual outcome. In other words, instead of directly identifying the optimal dose for each individual, SSITS tries to find the optimal regime that can result in a higher expected counterfactual outcome for the whole population, which makes a large part of the population benefit from the estimated optimal dosage

regime.

We further set $k = 20$, thus a=0.71 and b=-3.07, to mimic a situation when the tailoring variables that interact with the dose assignment are not clear to clinicians. In this setting, we intend to see how the noise interference may affect the performance of our proposed method. The results are summarized in Table 4.3. The comparison with existing methods shows similar trends as in Table 4.2. SSITS clearly outperforms the existing methods in most cases, especially in the cases with tree-type underlying dosage regimes. However, SSITS is more sensitive to the increase of covariates dimension, compared to the existing methods. In particular, CZK performs competitively in Scenario 5 when N=300, which is a complex non-tree-type underlying dosage regime where the pre-specified optimal dose of each patient cannot be classified to limited categories. The simulation studies from *Chen et al.* (2016) also demonstrated its capability of dealing with some non-linear dosage regime cases, especially with a small sample size. Our findings here are consistent with their results. The compromised performance of SSITS may derive from the lower convergence rate of the non-parametric method we used in the dose-response model. This is an issue that we plan on investigating more in the future.

4.5.2 Two-Stage Scenarios

In the setting of dynamic dosage regimes where the treatment decisions are made at multiple stages, SSITS can be utilized sequentially. Specifically, to apply in a two-stage scenario, suppose the outcome of interest is the sum of intermediate outcomes at each stage, i.e., $Y = Y_1 + Y_2$. d

Like the single stage scenarios, we also consider the underlying true optimal dosage regime as tree-type or non-tree-type. The data are generated by extending the parameter specifications of the single-stage scenarios (Scenario 1 and 4 from the single-stage scenarios). To help make decisions in the second stage, we generate two

Table 4.2: Simulation results for single-stage scenarios that use 5 baseline covariates (100 replications, N=300 or 500). 7 scenarios belong to two types of pre-specified underlying dosage regime: tree-type dosage regime (I) and non-tree-type dosage regime(II). $E\{Y^*(\widehat{g}^{opt})\} = 10$

	Scenario	Method	N=300		N=500	
			$E\{Y^*(\widehat{g}^{opt})\}$	$\mid d^{opt} - \widehat{d}^{opt} \mid$	$E\{Y^*(\widehat{g}^{opt})\}$	$\mid d^{opt} - \widehat{d}^{opt} \mid$
I	1	LZ	5.86(5.65, 6.34)	0.12(0.10, 0.13)	6.10(5.90, 6.53)	0.11(0.09, 0.11)
		CZK	6.44(6.28, 6.62)	0.14(0.13, 0.15)	6.61(6.47, 6.72)	0.13(0.12, 0.13)
		SSITS	9.70(9.56, 9.92)	0.02(0.01, 0.03)	9.83(9.65, 9.92)	0.01(0.01, 0.02)
	2	LZ	6.32(6.04, 6.65)	0.13(0.11, 0.14)	6.63(6.40, 7.20)	0.11(0.09, 0.12)
		CZK	6.34(6.18, 6.46)	0.14(0.13, 0.14)	6.48(6.36, 6.56)	0.13(0.12, 0.13)
		SSITS	8.11(7.56, 8.52)	0.10(0.07, 0.12)	9.10(8.37, 9.36)	0.05(0.03, 0.09)
	3	LZ	6.13(5.78, 6.55)	0.12(0.11, 0.14)	6.64(6.35, 6.96)	0.10(0.09, 0.11)
		CZK	6.56(6.44, 6.66)	0.12(0.12, 0.13)	6.64(6.52, 6.76)	0.12(0.11, 0.12)
		SSITS	7.97(7.25, 8.53)	0.08(0.06, 0.11)	9.54(9.41, 9.68)	0.02(0.01, 0.03)
II	4	LZ	5.64(5.46, 5.85)	0.14(0.13, 0.16)	5.77(5.62, 6.01)	0.13(0.12, 0.14)
		CZK	6.55(6.40, 6.65)	0.13(0.12, 0.14)	6.63(6.50, 6.77)	0.13(0.12, 0.13)
		SSITS	9.20(9.06, 9.28)	0.05(0.04, 0.06)	9.23(9.12, 9.31)	0.05(0.04, 0.05)
	5	LZ	6.46(6.18, 6.79)	0.13(0.11, 0.14)	7.11(6.87, 7.32)	0.10(0.09, 0.11)
		CZK	7.13(6.88, 7.31)	0.10(0.09, 0.11)	7.26(7.11, 7.37)	0.09(0.08, 0.09)
		SSITS	7.22(6.74, 7.61)	0.10(0.08, 0.13)	7.55(7.33, 7.82)	0.09(0.08, 0.10)
	6	LZ	6.03(5.69, 6.40)	0.17(0.14, 0.18)	6.42(6.17, 6.77)	0.14(0.13, 0.15)
		CZK	6.55(6.41, 6.72)	0.12(0.11, 0.13)	6.71(6.59, 6.84)	0.12(0.11, 0.12)
		SSITS	7.24(6.82, 7.62)	0.13(0.11, 0.16)	8.63(8.46, 8.80)	0.08(0.07, 0.08)
	7	LZ	6.44(6.07, 6.74)	0.16(0.14, 0.18)	6.90(6.58, 7.12)	0.14(0.13, 0.15)
		CZK	6.39(6.26, 6.51)	0.13(0.13, 0.14)	6.55(6.36, 6.67)	0.12(0.12, 0.13)
		SSITS	7.34(6.98, 7.84)	0.13(0.12, 0.17)	7.64(7.13, 8.15)	0.12(0.10, 0.15)

a. $\mid d^{opt} - \widehat{d}^{opt} \mid$ shows the median and its interquartile range of the difference between the true optimal dose and the estimated optimal dose.
b. $\widehat{E}\{Y^*(\widehat{g}^{opt})\}$ shows the median and the interquartile range of the estimated mean counterfactual outcome obtained using the true outcome model and the estimated optimal dynamic dosage regime.

Table 4.3: Simulation results for single-stage scenarios that use 20 baseline covariates (100 replications, N=300 or 500). 7 scenarios belong to two types of pre-specified underlying dosage regime: tree-type dosage regime (I) and non-tree-type dosage regime(II). $E\{Y^*(\widehat{g}^{opt})\} = 10$

	Scenario	Method	N=300		N=500	
			$E\{Y^*(\widehat{g}^{opt})\}$	$\mid d^{opt} - \widehat{d}^{opt} \mid$	$E\{Y^*(\widehat{g}^{opt})\}$	$\mid d^{opt} - \widehat{d}^{opt} \mid$
I	1	LZ	5.69(5.50, 6.08)	0.13(0.12, 0.15)	5.96(5.76, 6.45)	0.11(0.10, 0.12)
		CZK	5.89(5.77, 6.01)	0.18(0.17, 0.19)	6.05(5.92, 6.16)	0.16(0.16, 0.17)
		SSITS	9.66(9.56, 9.76)	0.02(0.02, 0.03)	9.70(9.65, 9.76)	0.02(0.02, 0.02)
	2	LZ	6.16(5.85, 6.43)	0.14(0.12, 0.16)	6.53(6.20, 6.78)	0.11(0.10, 0.13)
		CZK	5.83(5.76, 5.94)	0.17(0.16, 0.18)	5.97(5.87, 6.06)	0.16(0.15, 0.17)
		SSITS	7.39(6.94, 8.04)	0.13(0.11, 0.16)	8.28(7.86, 8.79)	0.09(0.06, 0.12)
	3	LZ	6.17(5.84, 6.51)	0.13(0.11, 0.15)	6.65(6.25, 6.99)	0.10(0.09, 0.11)
		CZK	6.00(5.76, 6.15)	0.16(0.15, 0.18)	6.16(6.05, 6.29)	0.14(0.14, 0.15)
		SSITS	7.73(6.88, 8.15)	0.09(0.07, 0.13)	9.17(8.46, 9.40)	0.04(0.03, 0.06)
II	4	LZ	5.47(5.34, 5.71)	0.15(0.14, 0.16)	5.70(5.56, 5.91)	0.14(0.13, 0.15)
		CZK	6.02(5.85, 6.11)	0.17(0.16, 0.18)	6.19(6.04, 6.31)	0.15(0.14, 0.16)
		SSITS	9.07(8.84, 9.26)	0.05(0.05, 0.07)	9.17(8.95, 9.28)	0.05(0.04, 0.06)
	5	LZ	6.36(6.04, 6.57)	0.14(0.12, 0.16)	6.95(6.72, 7.19)	0.10(0.09, 0.12)
		CZK	6.32(6.15, 6.45)	0.14(0.13, 0.15)	6.63(6.48, 6.77)	0.12(0.11, 0.12)
		SSITS	6.02(5.57, 6.58)	0.18(0.14, 0.22)	6.86(6.41, 7.16)	0.11(0.10, 0.14)
	6	LZ	5.94(5.63, 6.27)	0.17(0.16, 0.19)	6.21(5.94, 6.49)	0.15(0.14, 0.17)
		CZK	5.97(5.85, 6.07)	0.16(0.15, 0.17)	6.19(6.06, 6.31)	0.14(0.14, 0.15)
		SSITS	6.67(6.17, 7.37)	0.16(0.12, 0.19)	7.84(7.54, 8.60)	0.10(0.08, 0.11)
	7	LZ	6.17(5.80, 6.53)	0.18(0.16, 0.20)	6.66(6.31, 6.87)	0.15(0.14, 0.17)
		CZK	5.92(5.78, 6.03)	0.17(0.16, 0.18)	6.12(6.01, 6.20)	0.15(0.15, 0.16)
		SSITS	6.53(6.13, 7.02)	0.20(0.15, 0.23)	6.78(6.31, 7.21)	0.18(0.15, 0.21)

a. $\mid d^{opt} - \widehat{d}^{opt} \mid$ shows the median and its interquartile range of the difference between the true optimal dose and the estimated optimal dose.
b. $\widehat{E}(Y^*(\widehat{g}^{opt}))$ shows the median and the interquartile range of the expected mean counterfactual outcome obtained using the true outcome model and the estimated optimal dynamic dosage regime.

time-varying biomarkers X_4 and X_5, simulated independently from $Uniform(0,1)$, in addition to three baseline covariates X_1, X_2, X_3. For the first stage, we generate D_1 from $Beta(\alpha, 1-\alpha)$ with $\alpha = 1/\{1+\exp(0.3X_1+0.2X_2+0.1X_3)\}$. The continuous outcome of the first stage Y_1 has the form of $Y_1 = m_1(D_1) + \beta\mathbf{H}_1 + \epsilon_1$, where $\beta\mathbf{H}_1 = 0.5X_1 + 0.3X_2 + 0.7X_3$ and $m_1(d_1) = 1.1 \times F_{24,24}(0.5 - (d_1^{opt} - d_1))$; and d_1^{opt} is the optimal dose under the optimal dosage regime g_1^{opt}. We assume the independent standard normal variate $\epsilon_1 \sim N(0,1)$. Specifically, the underlying optimal dosage regime of stage 1 for the tree-type dosage regime is specified as $g_1^{opt}(\mathbf{H}_1) = 0.8I(X_1 > 0.7) + 0.2I(X_1 \leq 0.7)I(X_2 \leq 0.5) + 0.5I(X_1 \leq 0.7)I(X_2 > 0.5)$.The optimal dosage regime for the non tree-type is specified as $g_1^{opt}(\mathbf{H}_1) = 0.8I(X_1+X_2 > 1.2) + 0.2I(X_1+X_2 \leq 1.2)I(X_2 \leq 0.5) + 0.5I(X_1+X_2 \leq 1.2)I(X_2 > 0.5)$. At the second stage, the treatment assignment D_2 is also generated from $Beta(\alpha, 1-\alpha)$ with $\alpha = 1/\{1+\exp(0.3X_5+0.2X_5+0.1X_3)\}$. The continuous outcome $Y_2 = m_2(D_2) + \beta\mathbf{H}_2 + \epsilon_2$, where $\beta\mathbf{H}_2 = 0.3X_1 + 0.2X_2 + 0.5X_3 + 0.6X_4 + 0.2X_5$ and $m_2(d_2) = 1.1 \times F_{24,24}(0.5 - (d_2^{opt} - d_2))$. Again, we assume $\epsilon_2 \sim N(0,1)$. The underlying optimal dosage regime of stage 2 for the tree-type dosage regime is specified as $g_2^{opt}(\mathbf{H}_2) = 0.8I(Y_1 > 4) + 0.2I(Y_1 \leq 4)I(X_4 \leq 0.5) + 0.5I(Y_1 \leq 4)I(X_4 > 0.5)$.The optimal dosage regime for the non tree-type is specified as $g_2^{opt}(\mathbf{H}_2) = 0.8I(Y_1+X_5 > 4.2) + 0.2I(Y_1+X_5 \leq 4.2)I(X_2 \leq 0.5) + 0.5I(Y_1+X_5 \leq 1.2)I(X_2 > 0.5)$.

Results for different sample sizes and underlying optimal dosage regimes are shown in Table 4.4. Given a tree-type underlying dynamic dosage regime with a sample size N=300, the median of the estimated mean counterfactual outcome of two stages of SSITS is 18.21, which is fairly close to the optimal value of 20. The median of the estimated mean counterfactual outcome is 8.45 for stage 1 and 9.94 for stage 2. The better performance of the second stage is as expected, because we adapt the backward induction to start the estimation from the final stage. In contrast, CZK only has 11.24 in the median of the estimated mean counterfactual outcome, and LZ

is 10.54. A similar superiority of SSITS is also found under non-tree-type underlying dynamic dosage regimes and a larger sample size N=500. Moreover, when the sample size increases, the performance of all methods is better. Specifically, the convergence rate of the dose-response function using non-parametric models is much slower than that of the parametric case; therefore, the relatively compromised performance with a smaller sample size is also to be expected. Moreover, when the underlying true dynamic dosage regime is misspecified, i.e., under the non-tree-type scenario, the performance of SSITS is still reasonably satisfactory, with $\widehat{E}(Y^*(\widehat{g}^{opt})) = 17.28$ when $N = 300$ and with $\widehat{E}(Y^*(\widehat{g}^{opt})) = 17.33$ when $N = 500$. It shows that SSITS is robust against mis-specification of the underlying true model.

4.6 Real Data Application: Estimating an Optimal Warfarin Dosage Regime

We illustrate the application of SSITS by an analysis of the data from *The International Warfarin Pharmacogenetics Consortium* (2009) (IWPC). Warfarin is a commonly used anticoagulant medicine worldwide, and appropriate Warfarin dosing is critical in clinical practice. In particular, overdosing patients poses a high risk of bleeding, while underdosing undermines Warfarin's treatment effect against thrombosis.

Identifying the optimal dose for Warfarin is still an open problem in the clinical community (*Kimmel et al.*, 2013; *Fredrikson et al.*, 2014). To predict the optimal Warfarin dose, IWPC proposed three linear regression models, which are modeled by clinical data, pharmacogenetics-clinical data, and fixed single-dose data, respectively. Such linear regression models are valid when most of the patients from the training dataset receive the optimal dose. However, later studies (*Holbrook et al.*, 2005) have found that these dosage rules might be suboptimal for patients from some certain

Table 4.4: Simulation results for two-stage scenarios that use three baseline covariates and two time-varying covariates (100 replications, N=500 or 300). Tree-type (I) and non tree-type (II) are two different pre-specified underlying optimal dosage regimes. $E\{Y^*(\widehat{g}^{opt})\} = 20$.

	Sample Size	Method	$\mid d_1^{opt} - \widehat{d}_1^{opt} \mid$	$\mid d_2^{opt} - \widehat{d}_2^{opt} \mid$	$E\{Y_1^*(\widehat{g}_1{}^{opt})\}$	$E\{Y_2^*(\widehat{g}_2{}^{opt})\}$	$E(Y^*\{\widehat{\mathbf{g}}^{opt})\}$
I	300	LZ	0.19(0.16, 0.22)	0.11(0.11, 0.13)	4.20(4.02, 4.72)	6.15(5.94, 6.52)	10.54(10.11, 10.99)
		CZK	0.15(0.13, 0.17)	0.14(0.13, 0.15)	4.97(4.66, 5.18)	6.33(6.14, 6.47)	11.24(10.87, 11.56)
		SSITS	0.02(0.01, 0.03)	0.00(0.00, 0.28)	8.45(8.22, 8.59)	9.94(6.08, 9.99)	18.21(12.77, 18.53)
	500	LZ	0.15(0.14, 0.18)	0.10(0.10, 0.11)	4.36(4.14, 5.14)	6.43(6.25, 6.74)	10.95(10.59, 11.64)
		CZK	0.14(0.12, 0.15)	0.13(0.12, 0.15)	5.08(4.79, 5.31)	6.47(6.27, 6.66)	11.47(11.23, 11.81)
		SSITS	0.01(0.00, 0.02)	0.00(0.00, 0.25)	8.52(8.39, 8.66)	9.94(6.53, 9.99)	18.35(14.46, 18.59)
II	300	LZ	0.19(0.18, 0.21)	0.14(0.13, 0.16)	4.32(4.14, 4.64)	6.06(5.92, 6.32)	10.45(10.19, 10.74)
		CZK	0.15(0.13, 0.17)	0.15(0.14, 0.16)	4.85(4.63, 5.19)	6.11(5.94, 6.25)	11.00(10.65, 11.30)
		SSITS	0.06(0.05, 0.09)	0.02(0.01, 0.05)	7.75(6.99, 7.94)	9.72(8.99, 9.86)	17.28(15.56, 17.67)
	500	LZ	0.17(0.15, 0.18)	0.14(0.13, 0.15)	4.42(4.25, 4.67)	6.27(6.11, 6.52)	10.76(10.42, 11.07)
		CZK	0.14(0.13, 0.15)	0.14(0.13, 0.15)	4.98(4.72, 5.28)	6.38(6.23, 6.54)	11.35(11.06, 11.70)
		SSITS	0.05(0.05, 0.06)	0.02(0.02, 0.26)	7.87(7.70, 7.96)	9.53(6.54, 9.66)	17.33(14.04, 17.59)

a. $\mid d_1^{opt} - \widehat{d}_1^{opt} \mid$ and $\mid d_2^{opt} - \widehat{d}_2^{opt} \mid$ show how close the estimated optimal dose is to the underlying true optimal dose of each stage.
b. $\widehat{E}\{Y_1^*(\widehat{g}_2^{opt})\}$, $\widehat{E}\{Y_2^*(\widehat{g}_2^{opt})\}$, and $\widehat{E}\{Y^*(\widehat{\mathbf{g}}^{opt})\}$ show the median and the interquartile range of the estimated mean counterfactual outcome obtained using the true outcome model and the estimated optimal dynamic dosage regime (and each stage separately).

subgroups, e.g., elderly patients.

Therefore, to improve the accuracy of existing dosage algorithms and give clinicians more insights into treating patients with heterogeneous characteristics, an individualized dosage strategy becomes an ideal solution. In the following analysis, we use SSITS to estimate the optimal Warfarin dose, and provide an interpretable optimal dosage regime. According to the previous dose prediction linear regression models, the pharmacogenetics-clinical model had better performance for predicting the optimal dose, compared with the two other models. Therefore, we included both pharmacogenetics and clinical variables into the analysis, such as age, gender, height, weight, CYP2C9 genotype, VKOOC1 genotype and the medication status of Cytochrome P450 enzyme and Amiodarone. Following *Wallace et al.* (2018); *Chen et al.* (2016), after excluding patients with missing values, there remained a total of 1498 patients with a weekly Warfarin dose ranging from 4.5 mg to 315 mg. In clinical practice, INR (the international normalized ratio), an index that measures the rapidity of blood clotting, is used to evaluate the treatment effect of Warfarin. For patients treated by Warfarin, the target INR is 2-3 and INR=2.5 is the ideal outcome. Thus, we converted the observed INR into a reward outcome as $2 - \sqrt{|\ 2.5 - \text{INR}\ |}$ for each patient; larger values were preferable.

We set hyper-parameters of SSITS based on potential clinical needs. In particular, we set the minimum node size=150. This is a setting where clinicians expect to see more tailoring variables in a complex tree-based dosage regime, since the number of patients in the terminal nodes is small.

The estimated dosage regime is shown in Figure 4.1. If a patient is an African American, it suggests a lower dose of 17.5 mg Warfarin per week. For the patients who are from other races and are older than 57, they should take 28 mg per week is recommended if they do not have VKOOC1 AG mutation; otherwise they are recommended to take 17.5 mg per week. For the patients who are from other races and

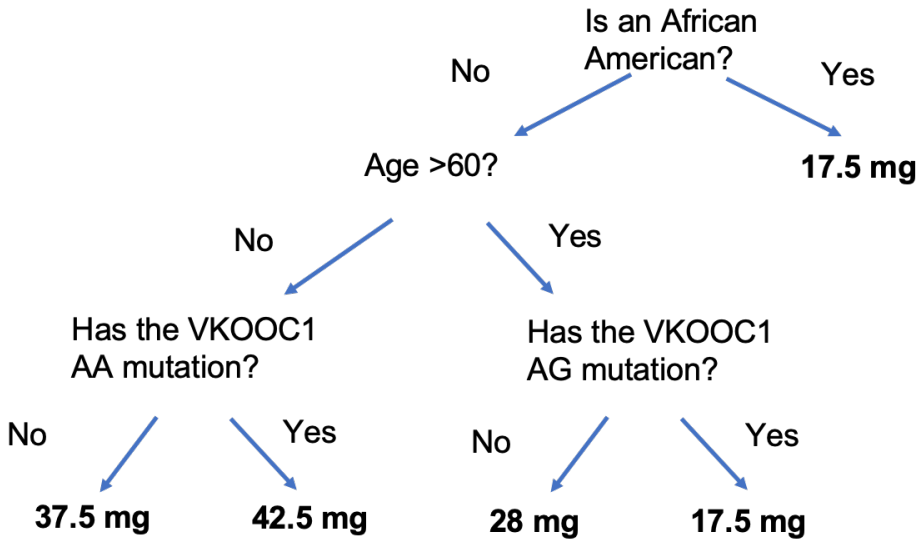

Figure 4.1: The estimated optimal Warfarin dosage regime estimated by SSITS using observational data from International Warfarin Pharmacogenetics Consortium.

younger than 60, the ones with VKOOC1 AA mutation, it is recommended to take 42.5 mg per week while the remaining patients should take 37.5 mg. These dose suggestions derived from the estimated dosage regime have consistent trends with what has been reported in the literature (*The International Warfarin Pharmacogenetics Consortium* , 2009; *Holbrook et al.*, 2005). In conclusion, SSITS is able to reveal some potential tailoring biomarkers that an optimal dose should be based on, and also shows its capability in estimating an individualized interpretable optimal dosage regime.

4.7 Discussion

Considering the limitations of current methods when estimating the optimal dynamic dosage regime, we developed the SSITS method to estimate the personalized optimal dosage regime. SSITS, which combines a robust non-parametric dose-response function with an efficient simulated annealing stochastic search method, has shown its consistently superior performance over existing methods in our comprehen-

sive simulation studies. In particular, even though the underlying optimal dosage regime of SSITS is not tree-type decision rule, SSITS still shows satisfactory performance when the underlying true optimal dosage regimes are mis-specified. In addition to the robustness, the easy-to-implement SSITS is able to provide interpretable decision rules. These make it particularly useful for clinicians to understand and apply the estimated regime with confidence. For a newly presenting patient, clinicians do not need to do extra complicated calculation and they are able to determine a recommended dose assignment right away.

According to the simulated annealing algorithm, the outcome of the iteration only depends on the outcome of the previous iteration, but the transition probabilities change with the number of iteration; therefore, the simulated annealing algorithm creates a time-inhomogeneous Markov chain. Under certain conditions, the discrete-time discrete-state Markov chain created by simulated annealing has been shown to converge to a stationary distribution of the optimal solution (*Lundy*, 1985; *Lundy and Mees*, 1986; *Granville et al.*, 1994). In practice, the convergence to global optimal requires excessive computational resource while the satisfactory empirical performance is straightforward to be obtained. One future research direction lies in further improving the computational efficiency of simulated annealing algorithm by incorporating methods from other discrete-space search approaches (*Osman*, 1995). Moreover, when increasing the number of covariates, the number of candidate decision trees have to be visited increases inevitably. Therefore, improving the algorithmic efficiency is also critical when applying this method in more complex dataset.

In addition, overfitting can be a problem in the procedure of estimation, since it is non-trivial to make inference for the estimated optimal tree-based decision rule. To overcome this issue, we may further incorporate the variance of dose response or other measure of uncertainty into the objective function $F(\cdot)$. That way, the tree with a high value in the objective function might be penalized for considerable uncertainty,

and thus the tree with the second or third highest value might be selected as the optimal one due to a more stable dose response.

Another possible direction is to extend our proposed method to other types of outcomes, such as binary outcomes, or time-to-event outcomes, as these outcomes are very common in clinical practice of oncology. How to balance the potentially competing or composite outcomes to determine the optimal dosage regime is also of great research interest. In clinical practice, clinicians may be interested in determining the optimal dosage regime according to multiple competing outcomes, such as overall survival, incidence of local progression, patient preference and quality of life. In order to balance these contradictory objectives, some statistical methods have been developed for multi-objective optimization, but none is able to estimate the optimal dosage on a continuum within the framework of dynamic treatment regime. One potential future direction is to further incorporate these multiple objective optimization functions and combine different objectives from different treatment stages, to estimate a practical and also viable optimal DTR.

CHAPTER V

Summary and Future Research Directions

In this dissertation, we have developed new statistical learning methods to evaluate dynamic treatment regimes, including estimating the test-and-treat decision rules using a new step-adjusted tree-based reinforcement learning method, and developing methods to estimate a robust and interpretable DTR with continuous dosage options.

The SAT-Learning method outlined in the second chapter is an important addition to the current DTR research on multi-stage and multi-treatment decision making. It provides health-care recommendations for testing and treating patients over time. We estimated the test-and-treat strategy by evaluating each test step and every treatment step embedded within each test step over multiple treatment stages. This step-adjusted DTR framework may have a greater impact with the rise of awareness of test screening when more screening data is available and more timely decisions regarding test-and-treat scheduling have to be made (*White et al.*, 2017; *Robertson and Ladabaum*, 2019).

In Chapter III and Chapter IV we studied two methods evaluating the dynamic treatment regimes with continuous treatment dosages. Assigning continuous doses over time is particularly important in treating patients with chronic diseases. By combining the DTR framework with the optimal individual dosing strategy, our methods SSITS and KIDD-Learning provide effective tools for physicians to offer more per-

sonalized doses over time. The stochastic tree search applied in SSITS allows the consideration of a richer number of candidate decision rules with an acceptable computational effort. In addition, the non-parametric method we used also relaxes the assumptions about the model of outcome and the structure of the candidate decision rule, and guarantees the flexibility of our methods.

We used the top-down tree-based reinforcement learning method to split the decision rule to compromise between exploration and exploitation in Chapter II and Chapter III; this method is improved in Chapter IV by stochastically walking within a broader space of decision trees. However, the number of decision rules will inevitably increase with the increasing number of tailoring variables, and the computational time will increase accordingly. One future research direction lies in improving the computational efficiency by incorporating other methods (*Cohen et al.*, 2003; *Qin et al.*, 2017) when applying our methods in more complex scenarios. Another research direction lies in further developing a variable selection method to screen potential tailoring variables for estimating the optimal DTR using observational datasets when high-dimension covariates exist (*Wallace et al.*, 2019; *Zhu et al.*, 2015).

Other extensions can also be considered to avoid the overfitting in the process of evaluating optimal DTR. Currently, the decision rule with the highest expected counterfactual outcome is chosen as the optimal rule. We might consider incorporating the uncertainty of this expected counterfactual outcome by penalizing some decision rules if they have high uncertainty, such as large variance. In this way, a decision rule with second or third highest expected counterfactual outcome might be acceptable due to the more stable performance. This can be extremely useful when the candidate treatment is on a continuous scale.

Another direction is to explore how to effectively balance the potentially contradictory outcomes and then to determine the optimal DTR. In clinical practice, the multiple competing outcomes may include incidence of local progression, patient pref-

erence, quality of life and cost burden (*Butler et al.*, 2017). In order to balance these contradictory objectives according to clinical needs, some statistical methods have been developed, but none has been generalized the framework of dynamic treatment regimes (*Lizotte and Laber*, 2016; *Laber et al.*, 2014). Incorporating these multiple objective optimization functions and aggregating different objectives from different treatment stages would be of great research interest in the future to estimate an optimal and also practical DTR.

APPENDICES

APPENDIX A

Stopping Rules for Chapter II

Algorithm 5 Stopping Rules

if the node size is less than $2n_0$ **then**

the node will not be split

end if

if all possible splits of a node result in a child node with size smaller than n_0 **then**

the node will not be split

end if

if the current tree depth reaches the user-specified maximum depth **then**

the node will not be split

end if

Calculate the best split by

$$\widehat{\omega}^{opt} = arg\max_{\omega} \left[\mathcal{P}_{sj}(\Omega, \omega) : \min\{n\mathbb{P}_n I(\mathbf{H}_{sj} \in \omega), n\mathbb{P}_n I(\mathbf{H}_{sj} \in \omega^c)\} \geq n_0\right].$$

if the maximum purity improvement $\mathcal{P}_{sj}(\Omega, \widehat{\omega}^{opt}) - \mathcal{P}_{sj}(\Omega) < \lambda$ **then**

the node will not be split

else

Split Ω into ω and ω^c

end if

APPENDIX B

Simulation Data Generating Process for Chapter II

Three covariates, X_1 to X_3, generated as baseline covariates follow $N(0,1)$. Two correlated covariates, X_4 and X_5, are generated as time-varying biomarkers which are measured just before the decision time of the test step within each stage. $(X_4, X_5)' \sim N(\mu, \Sigma)$, where $\mu = (0,0)'$ and $\Sigma = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}$. After the test step of each stage, the covariates X_{12} and X_{22} mimic the test results that contribute to the treatment decision nested within each test decision with other covariates. Typically, the test results, such as biopsy results, are of great importance to the treatment decision making. X_{12} and X_{22} follow the distribution of $N(0,1)$. To make the rates of taking the curative treatment equal to 5%, 15%, 20% and 25% in both stages, we also modify the parameters in the data generating models. More details of parameter setting are as follows:

Data Generation for Stage 1 The test decision variables, i.e., D_{11} and D_{21} are set to be the values of $\{0,1\}$ at the first step of each stage. For stage 1 step 1, we generate D_{11} from a $Bernoulli(\pi_{11,1})$ distribution with $\pi_{11,1} = \exp(0.2X_3 + X_4 - 0.5)/(1 + \exp(0.2X_3 + X_4 - 0.5))$. The reward of the stage 1 step 1 is generated as

$Y_{11} = X_4^2 + (0.5X_3 + 4)^2 \times I[g_{11}^{opt}(\mathbf{H}_{11}) = D_{11}] - 3|X_1|I(D_{11} = 1) + \epsilon_{11}$ with optimal regimes as

$$g_{11}^{opt}(\mathbf{H}_{11}) = I(X_1 > 0.3)I(X_4 \leq 1.3)$$

and $\epsilon_{11} \sim N(0, 1)$.

For patients who have been assigned the test, i.e., $D_{11} = 1$, we further generate the treatment assignment D_{12} for them as $D_{12} \sim Bernoulli(\pi_{12,1})$ with different treatment rates

$$\pi_{12,1} = \begin{cases} \exp(0.5X_{12} - X_2 - 3.3)/(1 + \exp(0.5X_{12} - X_2 - 3.3)) & \text{for rate=5\%} \\ \exp(0.5X_{12} - X_2 - 2.3)/(1 + \exp(0.5X_{12} - X_2 - 2.3)) & \text{for rate=15\%} \\ \exp(0.5X_{12} - X_2 - 1.8)/(1 + \exp(0.5X_{12} - X_2 - 1.8)) & \text{fort rate=20\%} \\ \exp(0.5X_{12} - X_2 - 1.5)/(1 + \exp(0.5X_{12} - X_2 - 1.5)) & \text{for rate=25\%} \end{cases}$$

We generate stage 1 step 2 reward as $Y_{12} = I[D_{12} = g_{12}^{opt}(\mathbf{H}_{12})](7+2X_4)+4X_3+Y_{11}/3+ 3I(D_{12} = 1)[I(g_{12}^{opt}(\mathbf{H}_{12}) = 1) - 1] + I(D_{12} = 1)(X_{12}^2 + 4) + \epsilon_{12}$ with $\epsilon_{12} \sim N(0, 1)$. The tree-type optimal regime at step 2 is specified as

$$g_{12}^{opt}(\mathbf{H}_{12}) = I(X_4 > 0.5)I(X_{12} \leq 0.3)$$

Data Generation for Stage 2: In stage 2, we generate the test decision $D_{21} \sim Bernoulli\,(\pi_{21,1})$ with $\pi_{21,1} = \exp(-0.7 + 0.1X_2 + X_5)/(1 + \exp(-0.7 + 0.1X_2 + X_5))$. The reward of stage 2 step 1 is generated as $Y_{21} = X_1^2 + 2X_2^2 + (8 - X_5)I[g_{21}^{opt}(\mathbf{H}_{21}) = D_{21}] - I(D_{21} = 1) + 4.5I(D_{21} = 1)[I(g_{21}^{opt}(\mathbf{H}_{21}) = 1) - 1] + \epsilon_{21}$ with $\epsilon_{21} \sim N(0, 1)$. The optimal regime is specified as

$$g_{21}^{opt}(\mathbf{H}_{21}) = I(X_2 < 0.8)I(X_5 > 0.1)$$

Among the patients who have had the test in the first step of stage 2, i.e., $D_{21} = 1$ we generate their treatment assignment D_{22} for the second step of the second stage. Specifically, we generate treatment $D_{22} \sim Bernoulli(\pi_{22,1})$ with different treatment rates

$$\pi_{22,1} = \begin{cases} \exp(0.5X_{22} - X_2 - 3.3)/(1 + \exp(0.5X_{22} - X_2 - 3.3)) & \text{for rate=5\%} \\ \exp(0.5X_{22} - X_2 - 2.3)/(1 + \exp(0.5X_{22} - X_2 - 2.3)) & \text{for rate=15\%} \\ \exp(0.5X_{22} - X_2 - 1.8)/(1 + \exp(0.5X_{22} - X_2 - 1.8)) & \text{for rate=20\%} \\ \exp(0.5X_{22} - X_2 - 1.5)/(1 + \exp(0.5X_{22} - X_2 - 1.5)) & \text{for rate=25\%} \end{cases}$$

The reward of stage 2 step 2 is generated as $Y_{22} = 3I[D_{22} = g_{22}^{opt}(\mathbf{H}_{22})](2X_{22} - X_5)^2 + Y_{21}/3 + (2X_4 + X_1) + \epsilon_{22}$ and $\epsilon_{22} \sim N(0, 1)$. The optimal treatment for stage 2 is specified as

$$g_{21}^{opt}(\mathbf{H}_{21}) = I(X_{22} < 0.3)I(X_5 > 0.5)$$

APPENDIX C

Data Preprocessing for Active Surveillance Data for Chapter II

For the prostate cancer data the exclusion criteria were the following: patients who did not have any PSA observations in the first 4 years were excluded and patients who were not followed after year 4 are excluded. For the remaining patients if they did not have a biopsy, the most recent PSA value that was used in the analysis was the last PSA within the time window between year 0 and year 2 for stage 1 and the last PSA value between year 2 and year 4 for stage 2. For patients who had a biopsy test, the most recent PSA for that test is the PSA value right before the date of biopsy. If a patient had more than one biopsy within a stage, we used the last biopsy result.

To assess the sensitivity of the estimated DTR tree in Figure 2.2 to modifications of the reward, we included an additional discounting factor for the reward of patients who had an especially high risk of future metastatic prostate cancer. Specifically, when a patient had his Gleason score ≥ 7 (4+3) during the first four years after diagnosis, his reward is reduced by a factor of 95%. The new estimated trees were very similar to the estimated optimal DTR shown in Figure 2.2, the only differences being small changes is the splitting thresholds at each node.

APPENDIX D

Generation Process of the Initial Tree $\mathcal{P}^1$ for Chapter IV

In a given iteration m, we assume $\mathcal{P}^m$ is sampled from $\pi(\mathcal{P})$. We further specify the tree topology model $\pi(\mathcal{P})$ using a stochastic tree-growing process, which can be decomposed as

$$\pi(\mathcal{P}) = \pi(\mathcal{T})\pi(\rho \mid \mathcal{T})\pi(\eta \mid \rho, \mathcal{T}),$$

where $\mathcal{T}$ is the tree topology, which includes the number of terminal nodes and how they are arranged; ρ denotes the splitting variables; η is the splitting thresholds.

A critical component of the specification of $\pi(\mathcal{P})$ is the specification of $\pi(\mathcal{T})$, which controls the number of terminal nodes and the skewness of the tree structure. A simple and intuitive way of growing a tree skeleton includes two steps: (1) sample the tree size, i.e., number of terminal nodes from a distribution; (2) cascade down these terminal nodes from the root like pouring pin balls. We draw the tree size $k(\mathcal{T})$ from a Poisson distribution $k(\mathcal{T}) \sim Pois(\lambda) + 1$. The 1 unit shift from the standard Poisson distribution avoids generating an empty tree, and the hyper parameter λ controls the complexity of the resulting tree structure. Moreover, we assume that an internal node u has $k_u(\mathcal{T})$) terminal nodes; each of these terminal nodes can be

assigned to the left or the right child node of u following a prespecified distribution. Thus, we can simplify the process as

$$\pi(\mathcal{T}) = \alpha(k(\mathcal{T})) \prod_{u \in a(\mathcal{T})} \beta(k_{u^*}(\mathcal{T}) \mid k_u(\mathcal{T})),$$

where $\alpha(\cdot)$ is the Poisson distribution that controls the tree size; $k(\mathcal{T})$ denotes the tree size. The conditional distribution $\beta(k_{u^*}(\mathcal{T}) \mid k_u(\mathcal{T}))$ dictates the shape of the tree, where $k_u(\mathcal{T}))$ is the number of available terminal nodes at internal node u and $k_{u^*}(\mathcal{T})$ is the number of nodes sent to the left child node u^* of u. The tree-growing process stops when $k_u(\mathcal{T}) = 1$ at any node u. In general, $\beta(\cdot)$ controls the preference of skewness of a tree and $\alpha(\cdot)$ governs the preference of the tree complexity. More details of the distribution specification of $\beta(\cdot)$ and $\alpha(\cdot)$ can be found in *Wu et al.* (2007).

For a given tree skeleton $\mathcal{T}$, the distributions of $\pi(\rho \mid \mathcal{T})$ and $\pi(\eta \mid \rho, \mathcal{T})$ are set to be uniform; i.e., each variable has equal probabilities of being selected and the splitting thresholds are determined uniformly within the domain of the selected variables.

BIBLIOGRAPHY

BIBLIOGRAPHY

Bertsimas, D., and J. Dunn (2017), Optimal classification trees, *Machine Learning*, *106*(7), 1039–1082.

Braun, T. M., S. Kang, and J. M. Taylor (2016), A Phase I/II trial design when response is unobserved in subjects with dose-limiting toxicity., *Statistical methods in medical research*, *25*(2), 659–73.

Breiman, L. (2001), Random forests, *Machine Learning*, *45*(1), 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984), *Classification and Regression Trees.*, Wadsworth, Belmont, CA.

Butler, E. L. (2016), Using Patient Preferences to estimate optimal treatment strategies for competing outcomes, Ph.D. thesis, The University of North Carolina at Chapel Hill.

Butler, E. L., E. B. Laber, S. M. Davis, and M. R. Kosorok (2017), Incorporating Patient Preferences into Estimation of Optimal Individualized Treatment Rules, *Biometrics*.

Chakraborty, B., and S. A. Murphy (2014), Dynamic Treatment Regimes, *Annual Review of Statistics and Its Application*, *1*(1), 447–464.

Chen, G., D. Zeng, and M. R. Kosorok (2016), Personalized Dose Finding Using Outcome Weighted Learning, *Journal of the American Statistical Association*, *111*(516), 1509–1521.

Chen, J., H. Fu, X. He, M. R. Kosorok, and Y. Liu (2018), Estimating individualized treatment rules for ordinal treatments, *Biometrics*, *74*(3), 924–933.

Chen, T., and C. Guestrin (2016), Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, ACM.

Cheung, Y. K. (2011), *Dose finding by the continual reassessment method*, 186 pp., CRC Press/Taylor & Francis Group.

Chipman, H. A., E. I. George, and R. E. Mcculloch (1998), Bayesian CART Model Search, *Journal of the American Statistical Association*, *93*, 935–948.

Cohen, M. B., C. J. Colbourn, and A. C. Ling (2003), Augmenting simulated annealing to build interaction test suites, in *14th International Symposium on Software Reliability Engineering, 2003. ISSRE 2003.*, pp. 394–405, IEEE.

Collins, F. S., and H. Varmus (2015), A new initiative on precision medicine, *New England journal of medicine*, *372*(9), 793–795.

Cui, Y., R. Zhu, and M. Kosorok (2017), Tree based weighted learning for estimating individualized treatment rules with censored data., *Electronic journal of statistics*, *11*(2), 3927–3953.

Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998), A Bayesian CART algorithm, *Biometrika*, *85*(2), 363–377.

Denton, B. T., S. T. Hawley, and T. M. Morgan (2019), Optimizing Prostate Cancer Surveillance: Using Data-driven Models for Informed Decision-making., *European Urology*, *75*(6), 918.

El Naqa, I., M. R. Kosorok, J. Jin, M. Mierzwa, and R. K. Ten Haken (2018), Prospects and Challenges for Clinical Decision Support in the Era of Big Data, *JCO Clinical Cancer Informatics*, *2*(2), 1–12.

Feng, M., et al. (2018), Individualized Adaptive Stereotactic Body Radiotherapy for Liver Tumors in Patients at High Risk for Liver Damage, *JAMA Oncology*, *4*(1), 40.

Feng, M. U., et al. (2013), A Phase 2 Trial of Individualized Adaptive Stereotactic Body Radiation Therapy (SBRT) for Patients at High Risk for Liver Damage, *International Journal of Radiation Oncology• Biology• Physics*, *87*(2), S27.

Flores, C. A., A. Flores-Lagunes, A. Gonzalez, and T. C. Neumann (2012), Estimating the effects of length of exposure to instruction in a training program: The case of job corps, *Review of Economics and Statistics*, *94*(1), 153–171.

Fredrikson, M., E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart (2014), Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing, in *23rd USENIX Security Symposium (USENIX Security 14)*, pp. 17–32.

Granville, V., M. Krivánek, and J.-P. Rasson (1994), Simulated annealing: A proof of convergence, *IEEE transactions on pattern analysis and machine intelligence*, *16*(6), 652–656.

Guo, B., and Y. Yuan (2017), Bayesian Phase I/II Biomarker-Based Dose Finding for Precision Medicine With Molecularly Targeted Agents, *Journal of the American Statistical Association*, *112*(518), 508–520.

Hanley, J. A. (2011), Measuring mortality reductions in cancer screening trials, *Epidemiologic reviews*, *33*(1), 36–45.

Hastie, T., and R. Tibshirani (1993), Varying-Coefficient Models, *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(4), 757–779.

Hayfield, T., and J. S. Racine (2008), Nonparametric econometrics: The np package, *Journal of statistical software*, *27*(5), 1–32.

Hirano, K., and G. W. Imbens (2005), The Propensity Score with Continuous Treatments, in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, pp. 73–84, John Wiley & Sons.

Holbrook, A. M., J. A. Pereira, R. Labiris, H. McDonald, J. D. Douketis, M. Crowther, and P. S. Wells (2005), Systematic overview of warfarin and its drug and food interactions, doi:10.1001/archinte.165.10.1095.

Huang, X., S. Choi, L. Wang, and P. F. Thall (2015), Optimization of multi-stage dynamic treatment regimes utilizing accumulated data, *Statistics in Medicine*, *34*(26), 3424–3443.

Imai, K., and D. A. Van Dyk (2004), Causal inference with general treatment regimes: Generalizing the propensity score, *Journal of the American Statistical Association*, *99*(467), 854–866.

Inoue, L. Y. T., D. W. Lin, L. F. Newcomb, A. S. Leonardson, D. Ankerst, R. Gulati, and Others (2018), Comparative Analysis of Biopsy Upgrading in Four Prostate Cancer Active Surveillance Cohorts, *Annals of Internal Medicine*, *168*(1), 1.

Ishwaran, H., and M. Lu (2019), Random Survival Forests, in *Wiley StatsRef: Statistics Reference Online*, pp. 1–13, John Wiley & Sons, Ltd, Chichester, UK.

Itzkowitz, S., R. Brand, L. Jandorf, K. Durkee, J. Millholland, L. Rabeneck, and Others (2008), A simplified, noninvasive stool DNA test for colorectal cancer detection, *The American Journal of Gastroenterology*, *103*(11), 2862.

Jackson, W., et al. (2019), A mid-treatment increase in albi score is strongly associated with treatment related toxicity following liver radiation therapy, *International Journal of Radiation Oncology• Biology• Physics*, *105*(1), S206–S207.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis (2004), kernlab-an S4 package for kernel methods in R, *Journal of statistical software*, *11*(9), 1–20.

Karatzoglou, A., A. Smola, K. Hornik, and M. A. Karatzoglou (2018), Package, *Tech. rep.*, Technical report, CRAN, 03 2016.

Kennedy, E. H., Z. Ma, M. D. Mchugh, and D. S. Small (2017), Non-parametric methods for doubly robust estimation of continuous treatment effects, *Journal of Royal Statistics Society. B*, *79*, 1229–1245.

Kimmel, S. E., et al. (2013), A pharmacogenetic versus a clinical algorithm for warfarin dosing, *New England Journal of Medicine*, *369*(24), 2283–2293.

Klotz, L., D. Vesprini, P. Sethukavalan, V. Jethava, L. Zhang, S. Jain, and Others (2014), Long-term follow-up of a large active surveillance cohort of patients with prostate cancer, *Journal of Clinical Oncology*, *33*(3), 272–277.

Laber, E. B., and Y. Q. Zhao (2015), Tree-based methods for individualized treatment regimes, *Biometrika*, *102*(3), 501–514.

Laber, E. B., D. J. Lizotte, and B. Ferguson (2014), Set-valued dynamic treatment regimes for competing outcomes, *Biometrics*, *70*(1), 53–61.

Lange, J. M., R. Gulati, A. S. Leonardson, D. W. Lin, L. F. Newcomb, B. J. Trock, and Others (2018), Estimating and comparing cancer progression risks under varying surveillance protocols, *The Annals of Applied Statistics*, *12*(3), 1773.

Laurent, H., and R. L. Rivest (1976), Constructing optimal binary decision trees is NP-complete, *Information processing letters*, *5*(1), 15–17.

Lawrence, T. S., J. M. Robertson, M. S. Anscher, R. L. Jirtle, W. D. Ensminger, and L. F. Fajardo (1995), Hepatic toxicity resulting from cancer treatment, *International Journal of Radiation Oncology• Biology• Physics*, *31*(5), 1237–1248.

Lee, J., P. F. Thall, Y. Ji, and P. Müller (2015), Bayesian Dose-Finding in Two Treatment Cycles Based on the Joint Utility of Efficacy and Toxicity, *Journal of the American Statistical Association*, *110*(510), 711–722.

Li, P., J. M. Taylor, S. Kong, S. Jolly, and M. J. Schipper (2019), A utility approach to individualized optimal dose selection using biomarkers, *Biometrical Journal*, p. bimj.201900030.

Lizotte, D. J., and E. B. Laber (2016), Multi-Objective Markov Decision Processes for Data-Driven Decision Support, *Journal of Machine Learning Research*, *17*, 1–28.

Lizotte, D. J., M. Bowling, and M. S. A (2012), Linear Fitted-Q Iteration with Multiple Reward Functions, *Journal of Machine Learning Research*, *13*, 3253–3295.

Loeb, S., M. A. Bjurlin, J. Nicholson, T. L. Tammela, D. F. Penson, H. B. Carter, and Others (2014), Overdiagnosis and overtreatment of prostate cancer, *European Urology*, *65*(6), 1046–1055.

Lundy, M. (1985), Applications of the annealing algorithm to combinatorial problems in statistics, *Biometrika*, *71*(1), 191–199.

Lundy, M., and A. Mees (1986), Convergence of an annealing algorithm, *Mathematical Programming*, *34*(1), 111–124.

Mandelblatt, J. S., K. A. Cronin, S. Bailey, D. A. Berry, H. J. De Koning, G. Draisma, and Others (2009), Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms, *Annals of internal medicine*, *151*(10), 738–747.

Mohammadi, H., et al. (2018), Using the albumin-bilirubin (albi) grade as a prognostic marker for radioembolization of hepatocellular carcinoma, *Journal of Gastrointestinal Oncology*, *9*(5), 840.

Morris, E., D. Owen, K. Cuneo, R. Ten Haken, M. Matuszak, T. Lawrence, and M. Schipper (2019), Albi and mean liver dose are predictive of overall survival in hepatocellular carcinoma patients treated with sbrt, *International Journal of Radiation Oncology• Biology• Physics*, *105*(1), S171–S172.

Moyer, V. A. (2012), Screening for Prostate Cancer: U.S. Preventive Services Task Force Recommendation Statement, *Annals of Internal Medicine*, *157*(2), 120.

Murphy, S. A. (2003), Optimal dynamic treatment regimes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 331–355.

Murphy, S. A., M. J. van der Laan, and J. M. Robins (2001), Marginal Mean Models for Dynamic Regimes, *Journal of the American Statistical Association*, *96*(456), 1410–1423.

Murthy, S., and S. Salzberg (1995), Lookahead and Pathology in Decision Tree Induction, in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pp. 1025–1031, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Niculescu, A. B., et al. (2019), Towards precision medicine for pain: diagnostic biomarkers and repurposed drugs, *Molecular Psychiatry*, p. 1.

Orellana, L., J. M. Robins, and A. Rotnitzky (2010), Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content "Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content Dynamic, *The International Journal of Biostatistics CAUSAL The International Journal of Biostatistics*, *6*(2).

Osman, I. H. (1995), Heuristics for the generalised assignment problem: simulated annealing and tabu search approaches, *Operations-Research-Spektrum*, *17*(4), 211–225.

Papagelis, A., and D. Kalles (2001), Breeding decision trees using evolutionary techniques, *Proceedings of the 18th International Conference on Machine Learning*, pp. 393–400.

Qin, L., J. Wang, H. Li, Y. Sun, and S. Li (2017), An approach to improve the performance of simulated annealing algorithm utilizing the variable universe adaptive fuzzy logic system, *IEEE Access*, *5*, 18,155–18,165.

Qu, Y., Z. Liu, H. Fu, S. Sethuraman, and P. M. Kulkarni (2019), Modeling the impact of preplanned dose titration on delayed response, *Journal of Biopharmaceutical Statistics*, *29*(2), 287–305.

Rich, B., E. E. Moodie, and D. A. Stephens (2016), Optimal individualized dosing strategies: A pharmacologic approach to developing dynamic treatment regimens for continuous-valued treatments, *Biometrical Journal*, *58*(3), 502–517.

Ridgeway, G., and M. G. Ridgeway (2004), The gbm package, *R Foundation for Statistical Computing, Vienna, Austria*, *5*(3).

Robertson, D. J., and U. Ladabaum (2019), Opportunities and challenges in moving from current guidelines to personalized colorectal cancer screening, *Gastroenterology*, *156*(4), 904–917.

Robins, J. (1986), A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect, *Mathematical Modelling*, *7*(9-12), 1393–1512.

Robins, J. M. (1989), The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies, *Health service research methodology: a focus on AIDS*, pp. 113–159.

Robins, J. M. (2004), Optimal structural nested models for optimal sequential decisions, in *Proceedings of the second seattle Symposium in Biostatistics*, pp. 189–326, Springer.

Robins, J. M., M. A. Hernan, and B. Brumback (2000), Marginal structural models and causal inference in epidemiology, *Epidemiology*, *11*, 550–560.

Rosenbaum, P. R. (1987), Model-based direct adjustment, *Journal of the American Statistical Association*, *82*(398), 387–394.

Salter, L. A., and D. K. Pearl (2001), Stochastic search strategy for estimation of maximum likelihood phylogenetic trees, *Systematic biology*, *50*(1), 7–17.

Scarpello, J. H. (2001), Review: Optimal dosing strategies for maximising the clinical response to metformin in type 2 diabetes, *The British Journal of Diabetes & Vascular Disease*, *1*(1), 28–36.

Schmidt, R. (1988), Dose-finding studies in clinical drug development, *European Journal of Clinical Pharmacology*, *34*(1), 15–19.

Schulte, P. J., A. A. Tsiatis, E. B. Laber, and M. Davidian (2014), Q- and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes, *Statistical Science*, *29*(4), 640–661.

Schulz, J., and E. E. M. Moodie (2020), Doubly Robust Estimation of Optimal Dosing Strategies, *Journal of the American Statistical Association*, pp. 1–31.

Shen, J., L. Wang, and J. M. G. Taylor (2017), Estimation of the optimal regime in treatment of prostate cancer recurrence from observational data using flexible weighting models, *Biometrics*, *73*(2), 635–645.

Shi, C., W. Lu, and R. Song (2020), Breaking the Curse of Nonregularity with Subagging Breaking the Curse of Nonregularity with Subagging-Inference of the Mean Outcome under Optimal Treatment Regimes, *Journal of Machine Learning Research*.

Sun, Y. (2019), Novel flexible statistical methods for missing data problems and personalized health care, Ph.D. thesis, University of Michigan, Ann Arbor.

Tao, Y., and L. Wang (2017), Adaptive contrast weighted learning for multi-stage multi-treatment decision-making, *Biometrics*, *73*(1), 145–155.

Tao, Y., L. Wang, and D. Almirall (2018), Tree-based reinforcement learning for estimating optimal dynamic treatment regimes, *The Annals of Applied Statistics*, *12*(3), 1914–1938.

Thall, P. F. (2008), Some geometric methods for constructing decision criteria based on two-dimensional parameters, *Journal of Statistical Planning and Inference*, *138*(2), 516–527.

Thall, P. F., and J. D. Cook (2004), Dose-Finding Based on Efficacy-Toxicity Trade-Offs, *Biometrics*, *60*(3), 684–693.

Thall, P. F., and K. E. Russell (1998), A Strategy for Dose-Finding and Safety Monitoring Based on Efficacy and Adverse Outcomes in Phase I/II Clinical Trials, *Biometrics*, *54*(1), 251.

Thall, P. F., L. H. Wooten, C. J. Logothetis, R. E. Millikan, and N. M. Tannir (2007), Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring, *Statistics in Medicine*, *26*(26), 4687–4702.

Thall, P. F., H. Q. Nguyen, and E. H. Estey (2008), Patient-specific dose finding based on bivariate outcomes and covariates, *Biometrics*, *64*(4), 1126–1136.

The International Warfarin Pharmacogenetics Consortium (2009), Estimation of the warfarin dose with clinical and pharmacogenetic data, *New England Journal of Medicine*, *360*(8), 753–764.

Tosoian, J. J., B. J. Trock, P. Landis, Z. Feng, J. I. Epstein, A. W. Partin, and Others (2011), Active Surveillance Program for Prostate Cancer: An Update of the Johns Hopkins Experience, *Journal of Clinical Oncology*, *29*(16), 2185–2190.

Tosoian, J. J., M. Mamawala, J. I. Epstein, P. Landis, S. Wolf, B. J. Trock, and H. B. Carter (2015), Intermediate and Longer-Term Outcomes From a Prospective Active-Surveillance Program for Favorable-Risk Prostate Cancer, *Journal of Clinical Oncology*, *33*(30), 3379–3385.

Trikalinos, T. A., U. Siebert, and J. Lau (2009), Decision-Analytic Modeling to Evaluate Benefits and Harms of Medical Tests: Uses and Limitations, *Medical Decision Making*, *29*(5), E22–E29.

US Preventive Services Task Force (2009), Screening for breast cancer: U.S.preventive services task force recommendation statement, *Annals of Internal Medicine*, *151*(10), 716.

Wallace, M. P., E. E. Moodie, and D. A. Stephens (2018), Reward ignorant modeling of dynamic treatment regimes, *Biometrical Journal*, *60*(5), 991–1002.

Wallace, M. P., E. E. Moodie, and D. A. Stephens (2019), Model selection for g-estimation of dynamic treatment regimes, *Biometrics*, *75*(4), 1205–1215.

Wang, L., A. Rotnitzky, X. Lin, R. E. Millikan, and P. F. Thall (2012), Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer, *Journal of the American Statistical Association*, *107*(498), 493–508.

Wang, T., C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille (2017), A Bayesian framework for learning rule sets for interpretable classification, *Journal of Machine Learning Research*, *18*, 1–37.

Wang, Y., Y. Zhao, and Y. Zheng (2019), Learning-based biomarker-assisted rules for optimized clinical benefit under a risk?constraint, *Biometrics*, p. biom.13199.

Watkins, C. J. C. H., and P. Dayan (1992), Q-learning, *Machine Learning*, *8*(3-4), 279–292.

White, A., T. D. Thompson, M. C. White, S. A. Sabatino, J. de Moor, P. V. Doria-Rose, A. M. Geiger, and L. C. Richardson (2017), Cancer screening test use—united states, 2015, *MMWR. Morbidity and mortality weekly report*, *66*(8), 201.

Wood, S. N. (2017), *Generalized additive models: an introduction with R*, CRC press.

Wu, Y., H. Tjelmeland, and M. West (2007), Bayesian CART: Prior Specification and Posterior Simulation, *Journal of Computational and Graphical Statistics*, *16*(1), 44–66.

Xu, Y., P. F. Thall, W. Hua, and B. S. Andersson (2018), Bayesian non?parametric survival regression for optimizing precision dosing of intravenous busulfan in allogeneic stem cell transplantation, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(3), rssc.12,331.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2013), Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions, *Biometrika*, *100*(3), 681–694.

Zhang, Y., E. B. Laber, A. Tsiatis, and M. Davidian (2015), Using decision lists to construct interpretable and parsimonious treatment regimes, *Biometrics*, *71*(4), 895–904.

Zhao, S., D. A. van Dyk, and K. Imai (2020), Propensity score-based methods for causal inference in observational studies with non-binary treatments, *Statistical Methods in Medical Research*, *29*(3), 709–727.

Zhao, Y. Q., D. Zeng, A. J. Rush, and M. R. Kosorok (2012), Estimating individualized treatment rules using outcome weighted learning, *Journal of the American Statistical Association*, *107*(499), 1106–1118.

Zhao, Y. Q., D. Zeng, E. B. Laber, and M. R. Kosorok (2015), New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes, *Journal of the American Statistical Association*, *110*(510), 583–598.

Zhu, R., and M. R. Kosorok (2012), Recursively imputed survival trees, *Journal of the American Statistical Association*, *107*(497), 331–340.

Zhu, R., D. Zeng, and M. R. Kosorok (2015), Reinforcement Learning Trees, *Journal of the American Statistical Association*, *110*(512), 1770–1784.